บทที่ 3

การวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุ

ในบทนี้จะกล่าวถึงการวิเคราะห์โมเดลถดถอยที่เกี่ยวข้องกับตัวแปรอิสระมากกว่าหนึ่งตัว ที่เรียกว่า *การวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุ* ซึ่งเป็นโมเดลที่ใช้กันอย่างแพร่หลาย และในทางปฏิบัติการสร้างสมการถดถอยแสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระเพียงหนึ่งตัวมักเกิดขึ้นไม่บ่อยนัก เช่น ในการสร้างสมการแสดงความสัมพันธ์ของยอดขาย นอกจากจะขึ้นอยู่กับค่าใช้จ่ายในการโฆษณาแล้ว ยังอาจมีตัวแปรอื่นที่อาจมีผลกระทบต่อยอดขาย เป็นต้นว่าจำนวนสินค้าที่ขาย ราคาขายต่อหน่วย และประสบการณ์ของพนักงานขาย โดยการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุสามารถทำได้โดยเขียนสมการให้อยู่ในรูปของเมตริกซ์ เพื่อให้สะดวกต่อการคำนวณ สำหรับการประมาณค่าพารามิเตอร์ในโมเดลถดถอย การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ การนำสมการถดถอยไปใช้เพื่อการพยากรณ์ และการหาช่วงความเชื่อมั่นของค่าพยากรณ์ สามารถทำได้ในลักษณะที่คล้ายคลึงกับการวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย

3.1 โมเดลถดถอยเชิงเส้นตรงแบบพหุ (Multiple Linear Regression Model)

โมเดลถดถอยเชิงเส้นตรงแบบพหุที่แสดงความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ k ตัว มีรูปแบบทั่วไปดังนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}, \qquad i = 1, 2, \ldots, n \tag{3.1}$$

เมื่อ

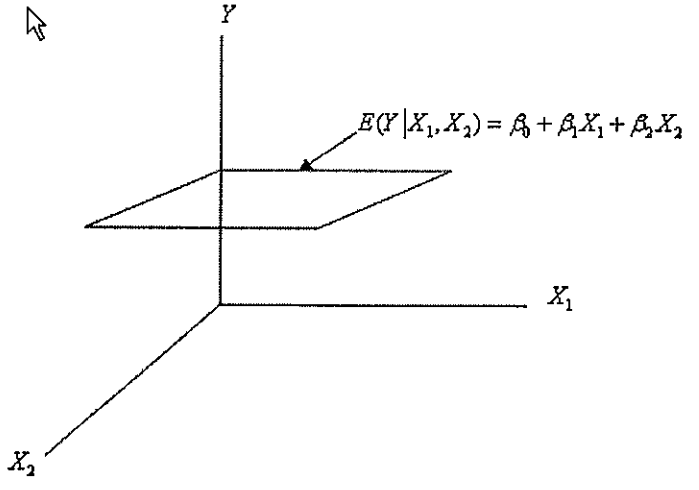
Y_i แทน ตัวแปรตาม

X_{ij} แทน ตัวแปรอิสระ เมื่อ $j = 0, \ldots, k$ และ $X_{i0} = 1$

β_j แทน ค่าสัมประสิทธิ์ถดถอย (Regression coefficients) เมื่อ $j = 0, \ldots, k$

ϵ_i แทน ความคลาดเคลื่อนสุ่ม (Random error)

โดยโมเดล (3.1) สามารถอธิบายได้ด้วยระนาบ (Hyperplane) ของตัวแปรอิสระ X_j ใน k มิติ และค่าสัมประสิทธิ์ถดถอย β_j เป็นพารามิเตอร์ที่ไม่ทราบค่า ซึ่งแทนการเปลี่ยนแปลงของค่าเฉลี่ยของ Y ต่อการเปลี่ยนแปลงของ X_j 1 หน่วย เมื่อให้ตัวแปรอิสระอื่นคงที่ ด้วยเหตุนี้จึงนิยมเรียกพารามิเตอร์ β_j เมื่อ $j = 1, 2, \ldots, k$ ว่า *ค่าสัมประสิทธิ์ถดถอยบางส่วน* (Partial regression coefficients)



รูปที่ 3.1: ระนาบของสมการถดถอยเชิงเส้นตรงแบบพหุเมื่อมีตัวแปรอิสระ 2 ตัว

พิจารณาโมเดลถดถอยเชิงเส้นตรงแบบพหุกรณี 2 ตัวแปร ซึ่งสามารถอธิบายได้ด้วยระนาบของตัวแปรอิสระ X_1 และ X_2 ใน 2 มิติ ดังแสดงในรูปที่ 3.1 จะเห็นได้ว่าค่า β_0 แทน ระยะตัดแกน Y ของระนาบถดถอย ถ้าขอบเขตของข้อมูลรวมค่า $X_1 = X_2 = 0$ แล้ว β_0 จะหมายถึงค่าเฉลี่ยของ Y เมื่อ $X_1 = X_2 = 0$ มิฉะนั้น β_0 จะไม่มีความหมาย ส่วนพารามิเตอร์ β_1 วัดการเปลี่ยนแปลงของค่าเฉลี่ยของ Y ต่อ 1 หน่วยการเปลี่ยน-แปลงของ X_1 เมื่อให้ X_2 คงที่ ทำนองเดียวกัน β_2 วัดการเปลี่ยนแปลงของค่าเฉลี่ยของ Y ต่อ 1 หน่วยการเปลี่ยนแปลงของ X_2 เมื่อให้ X_1 คงที่

บ่อยครั้งโมเดลถดถอยเชิงเส้นตรงแบบพหุจะถูกใช้เพื่อประมาณความสัมพันธ์ที่แท้จริงระหว่าง Y กับ $X_1, X_2, \ldots, X_k$ ที่ไม่ทราบรูปแบบที่แน่นอน ทราบแต่เพียงว่ารูปแบบเชิงเส้นตรงนั้นเหมาะสมภายในขอบเขตของข้อมูลที่ใช้ในการสร้างสมการโดยประมาณ นอกจากนี้โมเดลที่มีโครงสร้างซับซ้อนยังสามารถวิเคราะห์โดยใช้เทคนิคการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุได้อีกด้วย เช่น โมเดลควอดราติก (Quadratic model) ดังต่อไปนี้

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

ถ้าให้ $X_1 = X$ และ $X_2 = X^2$ แล้ว โมเดลข้างต้นสามารถเขียนใหม่ได้เป็น

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

ซึ่งก็คือรูปแบบโมเดลถดถอยเชิงเส้นตรงแบบพหุของ 2 ตัวแปรนั่นเอง โดยทั่วไปแล้วโมเดลถดถอยที่มีลักษณะเชิงเส้นตรงในเทอมของพารามิเตอร์ จะจัดเป็นโมเดลถดถอยเชิงเส้นตรงทั้งสิ้น โดยไม่คำนึงถึงลักษณะรูปร่างของพื้นผิวที่ได้จากโมเดลนั้น ๆ

ข้อตกลงเบื้องต้น (Assumptions) ของการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุ

- $\epsilon_i \sim NID\left(0,\, \sigma^2\right)$
- $Y_i \sim NID\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik},\, \sigma^2\right)$
- X_j เป็นตัวแปรที่ทราบค่า (Known value) เมื่อ $j = 1, 2, \ldots, k$
- X_j และ X_ℓ เป็นอิสระกัน เมื่อ $j,\, \ell = 1, 2, \ldots, k$

3.2 การประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุด

การประมาณค่าพารามิเตอร์ในโมเดลถดถอยเชิงเส้นตรงแบบพหุสามารถทำได้โดยใช้วิธีกำลังสองน้อยที่สุดเช่นเดียวกับการประมาณค่าพารามิเตอร์ในโมเดลถดถอยเชิงเส้นตรงอย่างง่าย จากการเก็บรวบรวมข้อมูลตัวอย่างซึ่งประกอบด้วยค่าสังเกต n ค่า เมื่อ $n > k$ ให้ Y_i แทน ค่าสังเกตตัวที่ i เมื่อ $i = 1, 2, \ldots, n$ และ X_{ij} แทน ค่าสังเกตตัวที่ i ของ X_j เมื่อ $j = 1, 2, \ldots, k$ จะได้ข้อมูลที่ใช้ในการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุดังนี้

ค่าสังเกตที่	Y_i	X_1	X_2	$\cdots$	X_k
1	Y_1	X_{11}	X_{12}	$\cdots$	X_{1k}
2	Y_2	X_{21}	X_{22}	$\cdots$	X_{2k}
3	Y_3	X_{31}	X_{32}	$\cdots$	X_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	Y_n	X_{n1}	X_{n2}	$\cdots$	X_{nk}

ตารางที่ 3.1: ลักษณะข้อมูลที่ใช้ในการวิเคราะห์สมการถดถอยเชิงเส้นตรงแบบพหุ

จากโมเดลถดถอยเชิงเส้นตรงแบบพหุ (3.1) สามารถเขียนใหม่ได้เป็น

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{3.2}$$

และได้ผลรวมกำลังสองของความคลาดเคลื่อนดังนี้

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{i=1}^{n} \beta_j X_{ij})^2 \end{aligned} \tag{3.3}$$

การหาตัวประมาณของ $\beta_0, \beta_1, \ldots, \beta_k$ ด้วยวิธีกำลังสองน้อยที่สุด ทำได้โดยหาอนุพันธ์ของ SSE เทียบกับพารามิเตอร์แต่ละตัวดังนี้

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{i=1}^{n} \beta_j X_{ij} \right) \tag{3.4}$$

$$\frac{\partial SSE}{\partial \beta_j} = -2 \sum_{i=1}^{n} X_{ij} \left(Y_i - \beta_0 - \sum_{i=1}^{n} \beta_j X_{ij} \right), \qquad j = 1, 2, \ldots, k \tag{3.5}$$

แล้วให้สมการ (3.4) และ (3.5) เท่ากับศูนย์ และแทนค่า $\beta_0, \beta_1, \ldots, \beta_k$ ด้วย $b_0, b_1, \ldots, b_k$ ตามลำดับ จะได้ว่า

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - b_0 - \sum_{i=1}^{n} b_j X_{ij} \right) &= 0 \\ \sum_{i=1}^{n} X_{ij} \left(Y_i - b_0 - \sum_{i=1}^{n} b_j X_{ij} \right) &= 0, \qquad j = 1, 2, \ldots, k \end{aligned}$$

หลังจากนั้นจัดรูปสมการข้างต้นใหม่ จะได้สมการปกติ $p = k + 1$ สมการ ดังนี้

$$
\begin{aligned}
nb_0 + b_1\sum_{i=1}^{n} X_{i1} + b_2\sum_{i=1}^{n} X_{i2} + \cdots + b_k\sum_{i=1}^{n} X_{ik} &= \sum_{i=1}^{n} Y_i \\
b_0\sum_{i=1}^{n} X_{i1} + b_1\sum_{i=1}^{n} X_{i1}^2 + b_2\sum_{i=1}^{n} X_{i1}X_{i2} + \cdots + b_k\sum_{i=1}^{n} X_{i1}X_{ik} &= \sum_{i=1}^{n} X_{i1}Y_i \\
b_0\sum_{i=1}^{n} X_{i2} + b_1\sum_{i=1}^{n} X_{i1}X_{i2} + b_2\sum_{i=1}^{n} X_{i2}^2 + \cdots + b_k\sum_{i=1}^{n} X_{i2}X_{ik} &= \sum_{i=1}^{n} X_{i2}Y_i \\
\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \quad &= \quad \vdots \\
b_0\sum_{i=1}^{n} X_{ik} + b_1\sum_{i=1}^{n} X_{i1}X_{ik} + b_2\sum_{i=1}^{n} X_{i2}X_{ik} + \cdots + b_k\sum_{i=1}^{n} X_{ik}^2 &= \sum_{i=1}^{n} X_{ik}Y_i \qquad (3.6)
\end{aligned}
$$

หลังจากแก้สมการปกติทั้ง $p = k + 1$ สมการ ก็จะได้ตัวประมาณกำลังสองน้อยที่สุดของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_k$ แต่การแก้สมการ $k+1$ สมการด้วยหลักพีชคณิตจะค่อนข้างยุ่งยาก โดยเฉพาะอย่างยิ่งเมื่อมีตัวแปรอิสระหลายตัว ดังนั้นเพื่อให้สะดวกต่อการคำนวณ สมการ (3.1) สามารถเขียนให้อยู่ในรูปของเมตริกซ์ได้ดังนี้

$$
\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (3.7)
$$

โดยที่ $\underset{(n\times 1)}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ 1 & X_{31} & X_{32} & \ldots & X_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix}$

$$
\underset{(p\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \underset{(n\times 1)}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

เมื่อ

$\mathbf{Y}$ แทน เวกเตอร์ของค่าสังเกตขนาด $n \times 1$

$\mathbf{X}$ แทน เมเตริกซ์ของค่าของตัวแปรอิสระขนาด $n \times p$, $p = k + 1$

$\boldsymbol{\beta}$ แทน เวกเตอร์ของค่าสัมประสิทธิ์ถดถอยขนาด $p \times 1$

$\boldsymbol{\epsilon}$ แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $n \times 1$

เขียนผลรวมกำลังสองให้อยู่ในรูปของเมตริกซ์ได้ดังนี้

$$
\begin{aligned}
SSE &= \sum_{i=1}^{n} \epsilon_i^2 \\
&= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\
&= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\
&= \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\
&= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \qquad (3.8)
\end{aligned}
$$

หาอนุพันธ์เทียบกับพารามิเตอร์ β แล้วให้สมการเท่ากับศูนย์ และแทนค่าพารามิเตอร์ด้วยตัวประมาณ $\mathbf{b}$

$$\left.\frac{\partial SSE}{\partial \beta}\right|_{\mathbf{b}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0 \qquad (3.9)$$

จะได้สมการปกติในรูปของเมตริกซ์เป็น

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y} \qquad (3.10)$$

แก้สมการ (3.10) โดยคูณ $(\mathbf{X}'\mathbf{X})^{-1}$ เข้าทั้งสองข้างของสมการ จะได้ตัวประมาณกำลังสองน้อยที่สุดของ β คือ

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \qquad (3.11)$$

เมื่อ $(\mathbf{X}'\mathbf{X})^{-1}$ หาค่าได้ โดยเมตริกซ์$(\mathbf{X}'\mathbf{X})^{-1}$ หาค่าได้เสมอถ้าตัวแปรอิสระไม่มีความสัมพันธ์กัน (Linearly independent) นั่นคือ ไม่มีคอลัมน์ใดของ X เป็นฟังก์ชันเชิงเส้นของคอลัมน์อื่น และสมการ (3.10) สามารถเขียนแสดงรายละเอียดได้ดังนี้

$$
\begin{bmatrix}
n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \dots & \sum_{i=1}^{n} X_{ik} \\
\sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \dots & \sum_{i=1}^{n} X_{i1}X_{ik} \\
\vdots & \vdots & \vdots & & \vdots \\
\sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \dots & \sum_{i=1}^{n} X_{ik}^2
\end{bmatrix}
\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} Y_i \\
\sum_{i=1}^{n} X_{i1}Y_i \\
\vdots \\
\sum_{i=1}^{n} X_{ik}Y_i
\end{bmatrix}
$$

จะเห็นได้ว่าเมตริกซ์ $\mathbf{X}'\mathbf{X}$ เป็นเมตริกซ์สมมาตร (Symmetric matarix) ที่มีรูปแบบเฉพาะ โดยค่าที่อยู่บนเส้นทแยงมุม (Diagonal elements) เป็นค่าผลรวมกำลังสองของคอลัมน์ของ $\mathbf{X}$ และค่านอกเส้นทแยงมุม (Off-diagonal elements) เป็นค่าผลรวมของผลคูณ (Cross product) ระหว่างคอลัมน์ของ $\mathbf{X}$ ส่วน $\mathbf{X}'\mathbf{Y}$

เป็นคอลัมน์เวกเตอร์ขนาด $p \times 1$ ที่เกิดจากผลรวมของผลคูณระหว่างคอลัมน์ของ $\mathbf{X}$ กับค่าสังเกต Y_i

ดังนั้นสมการถดถอยในรูปของเมตริกซ์ คือ

$$\hat{\mathbf{Y}} = \mathbf{Xb} \tag{3.12}$$

นอกจากนี้เวกเตอร์ของค่าพยากรณ์ $\hat{\mathbf{Y}}$ ยังมีความสัมพันธ์กับเวกเตอร์ของค่าสังเกต $\mathbf{Y}$ ดังนี้

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{Xb} \\ &= \mathbf{X(X'X)^{-1}X'Y} \\ &= \mathbf{HY}\end{aligned} \tag{3.13}$$

เมื่อ $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$ และเรียกว่า *Hat matrix* มีขนาด $n \times n$ ซึ่งเป็นการเชื่อมโยงค่าสังเกตไปยังค่าพยากรณ์

ทำนองเดียวกันเวกเตอร์ของความคลาดเคลื่อนสามารถเขียนในรูปของเมตริกซ์ได้เป็น

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \hat{\mathbf{X}}\mathbf{b} \\ &= \mathbf{Y} - \mathbf{HY} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned} \tag{3.14}$$

ตัวอย่างที่ 3.1 ในการศึกษาความสัมพันธ์ระหว่างความพึงพอใจในการทำงานของพนักงานในโรงพยาบาลแห่งหนึ่งว่าขึ้นอยู่กับระดับของสติปัญญาและการปรับตัวหรือไม่ จึงได้สุ่มพนักงานในโรงพยาบาลแห่งนี้มา 15 คน เก็บรวมรวมข้อมูลดังแสดงในตารางที่ 3.2 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุแสดงความสัมพันธ์ระหว่างความพึงพอใจกับระดับสติปัญญาและดัชนี การปรับตัว

วิธีทำ

จาก Matrix plot ในรูปที่ 3.2 จะเห็นได้ว่าความสัมพันธ์ระหว่างความพึงพอใจกับระดับสติปัญญาและการปรับตัวมีแนวโน้มเป็นเส้นตรง โดยมีความสัมพันธ์ไปในทิศทางดียวกัน นั่นคือ ถ้าคะแนนระดับสติปัญญาและดัชนีการ-ปรับตัวสูงขึ้น พนักงานมีความพึงพอใจในการทำงานมากขึ้นตาม

พนักงานคนที่	คะแนนความพึงพอใจ (Y)	คะแนนระดับสติปัญญา (X_1)	ดัชนีการปรับตัว (X_2)
1	56	17	9
2	39	15	2
3	32	17	2
4	50	17	8
5	39	12	5
6	39	16	3
7	33	10	4
8	51	14	8
9	45	3	10
10	14	5	2
11	32	17	2
12	39	16	3
13	63	16	10
14	33	11	2
15	33	6	6

ตารางที่ 3.2: ข้อมูลคะแนนความพึงพอใจ คะแนนระดับสติปัญญา และดัชนีการปรับตัว

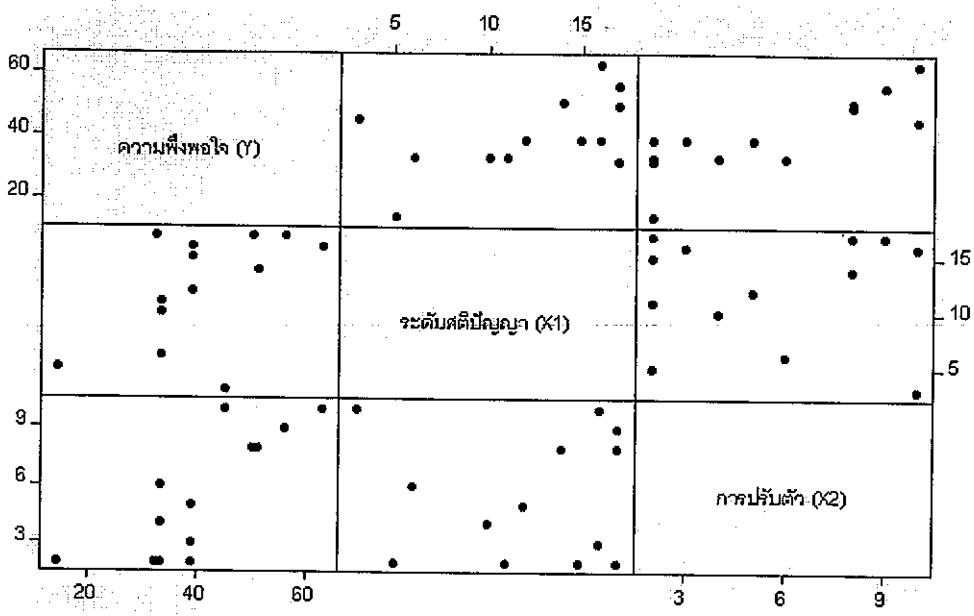
คำนวณค่าต่าง ๆ ของข้อมูลในตารางที่ 3.2

$$\sum_{i=1}^{n} X_{i1} = 192, \qquad \sum_{i=1}^{n} X_{i2} = 76, \qquad \sum_{i=1}^{n} Y_i = 598,$$
$$\sum_{i=1}^{n} X_{i1}^2 = 2,780, \qquad \sum_{i=1}^{n} X_{i2}^2 = 524, \qquad \sum_{i=1}^{n} X_{i1}X_{i2} = 953,$$
$$\sum_{i=1}^{n} X_{i1}Y_i = 8,009, \qquad \sum_{i=1}^{n} X_{i2}Y_i = 3,451, \qquad n = 15,$$
$$\bar{X}_1 = 12.8, \qquad \bar{X}_2 = 5.0667, \qquad \bar{Y} = 39.8667$$

แทนค่าลงในเมตริกซ์ได้เป็น

$$\mathbf{X'X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} \\ \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i2}^2 \end{bmatrix} = \begin{bmatrix} 15 & 192 & 76 \\ 192 & 2,780 & 953 \\ 76 & 953 & 524 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix}$$



รูปที่ 3.2: Matrix plot ของคะแนนความพึงพอใจ คะแนนระดับสติปัญญา และดัชนีการปรับตัว

$$\mathbf{X'Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \end{bmatrix} = \begin{bmatrix} 598 \\ 8,009 \\ 3,451 \end{bmatrix}$$

และได้ $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix}$

ดังนั้นสมการถดถอย คือ

$$\hat{Y} = 6.9652 + 1.2974X_1 + 3.2161X_2$$

นั่นคือ เมื่อให้ดัชนีการปรับตัวคงที่ ถ้าคะแนนระดับสติปัญญาสูงขึ้น 1 คะแนน ระดับความพึงพอใจจะเพิ่มขึ้น 1.2974 คะแนน และเมื่อให้ระดับสติปัญญาคงที่ ถ้าดัชนีการปรับตัวสูงขึ้น 1 คะแนน ระดับความพึงพอใจจะเพิ่มขึ้น 3.2161 คะแนน

3.2.1 คุณสมบัติของตัวประมาณแบบกำลังสองน้อยที่สุด

ตัวประมาณกำลังสองน้อยที่สุดมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียง (Unbiased estimator) ของพารามิเตอร์ β พิจารณาจาก

$$\begin{aligned}
E(\mathbf{b}) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left(\mathbf{X}\beta + \epsilon\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E\left(\beta\right) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E(\epsilon) \\
&= \beta
\end{aligned}$$

เนื่องจาก $E(\epsilon) = \mathbf{0}$ และ $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ ดังนั้น $\mathbf{b}$ เป็นตัวประมาณที่ไม่เอนเอียงของ β

พิจารณาเมตริกซ์ความแปรปรวน (Covariance matrix) ของตัวประมาณกำลังสองน้อยที่สุด

$$\begin{aligned}
Cov(\mathbf{b}) &= Cov\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] \\
&= Cov(\mathbf{AY}) \qquad\qquad \text{เมื่อ} \quad \mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{A}\, Cov(\mathbf{Y})\, \mathbf{A}' \\
&= \sigma^2\, \mathbf{AA}' \\
&= \sigma^2\, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2\, (\mathbf{X}'\mathbf{X})^{-1} \qquad (3.15)
\end{aligned}$$

โดย $Cov(\mathbf{b})$ เป็นเมตริกซ์สมมาตรขนาด $p \times p$ ซึ่งสมาชิกที่อยู่บนเส้นทแยงมุมเป็นความแปรปรวนของสัมประสิทธิ์ถดถอยแต่ละค่า และค่าที่อยู่นอกเส้นทแยงมุมเป็นความแปรปรวนร่วมของสัมประสิทธิ์ถดถอยแต่ละคู่ นั่นคือ ถ้าให้

$$\begin{aligned}
\mathbf{C} &= (\mathbf{X}'\mathbf{X})^{-1} \\
&= \begin{bmatrix} C_{00} & C_{01} & C_{02} & \dots & C_{0k} \\ C_{10} & C_{11} & C_{12} & \dots & C_{1k} \\ \vdots & \vdots & \vdots & & \vdots \\ C_{k0} & C_{k1} & C_{k2} & \dots & C_{kk} \end{bmatrix}
\end{aligned}$$

แล้วจะได้ว่า

$$
\begin{aligned}
V(b_j) &= \sigma^2 C_{jj}, \qquad j = 1, 2, \ldots, k \\
Cov(b_\ell, b_j) &= \sigma^2 C_{\ell j}
\end{aligned}
$$

เมื่อ C_{ij} แทน สมาชิกในตำแหน่งที่ (ij) ของเมตริกซ์ $\mathbf{C}$

ทำนองเดียวกับการวิเคราะห์การถดถอยอย่างง่าย พบว่าตัวประมาณกำลังสองน้อยที่สุดมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียงเชิงเส้นที่ดีที่สุด (BLUE) ของพารามิเตอร์ β และถ้าข้อมูลมีการแจกแจงแบบปกติแล้ว จะได้ว่าตัวประมาณกำลังสองน้อยที่สุดก็คือ ตัวประมาณแบบภาวะน่าจะเป็นสูงสุด (Maximum likelihood estimator: MLE) นั่นเอง

3.2.2 การประมาณค่าความแปรปรวนของความคลาดเคลื่อน σ^2

การประมาณค่าความแปรปรวนของความคลาดเคลื่อนจะคำนวณจากผลรวมกำลังสองของความคลาดเคลื่อนที่ได้จากข้อมูลตัวอย่าง ดังนี้

$$
\begin{aligned}
SSE &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} e_i^2 \\
&= \mathbf{e}'\mathbf{e}
\end{aligned}
$$

แทนค่า $\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$ ลงในสมการข้างต้น จะได้

$$
\begin{aligned}
SSE &= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\
&= \mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} \qquad (3.16)
\end{aligned}
$$

เนื่องจาก $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$

ค่าผลรวมกำลังสองของความคลาดเคลื่อนมีจำนวนองศาความเป็นอิสระที่สอดคล้องกันเป็น $n-p$ เมื่อ p แทน จำนวนพารามิเตอร์ที่ต้องประมาณในโมเดลถดถอย และได้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเป็น

$$MSE = S^2 = \frac{SSE}{n-p} = \frac{\mathbf{Y'Y} - \mathbf{b'X'Y}}{n-p} = \frac{\mathbf{Y'Y} - \mathbf{b'X'Y}}{n-k-1} \tag{3.17}$$

ดังนั้นตัวประมาณแบบจุดของ σ^2 คือ MSE หรือ S^2 ซึ่งมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียงของ σ^2 นั่นคือ

$$E(MSE) = \sigma^2$$

และเรียกรากที่สองของ MSE หรือ $\sqrt{MSE}$ ว่าเป็นค่าความคลาดเคลื่อนมาตรฐานของการประมาณ โดย MSE วัดความผันแปรรอบเส้นถดถอย จึงขึ้นอยู่กับรูปแบบของโมเดลเช่นเดียวกับตัวประมาณของ σ^2 ในการวิเคราะห์การถดถอยอย่างง่าย

ตัวอย่างที่ 3.2 จงคำนวณค่าประมาณของความแปรปรวนของความคลาดเคลื่อนในสมการถดถอยที่ได้ในตัวอย่างที่ 3.1

วิธีทำ คำนวณค่าต่าง ๆ ได้ดังนี้

$$\begin{aligned}
\mathbf{Y'Y} &= \sum_{i=1}^{n} Y_i^2 = 25,826 \\
\mathbf{b'X'Y} &= \begin{bmatrix} 6.9652 & 1.2974 & 3.2161 \end{bmatrix} \begin{bmatrix} 598 \\ 8,009 \\ 3,451 \end{bmatrix} = 25,654.7
\end{aligned}$$

คำนวณหาค่าผลรวมกำลังสองของความคลาดเคลื่อนได้ดังนี้

$$\begin{aligned}
SSE &= \mathbf{Y'Y} - \mathbf{b'X'Y} \\
&= 25,826 - 25,654.7 \\
&= 171.3
\end{aligned}$$

ดังนั้นค่าประมาณของ σ^2 คือ

$$
\begin{aligned}
MSE &= \frac{\mathbf{Y'Y} - \mathbf{b'X'Y}}{n-k-1} \\
&= \frac{25,826 - 25,654.7}{15-2-1} \\
&= 14.274
\end{aligned}
$$

3.3 การอนุมานทางสถิติเกี่ยวกับพารามิเตอร์ β

การอนุมานทางสถิติสำหรับพารามิเตอร์ β ต้องอาศัยข้อกำหนดเบื้องต้นเกี่ยวกับการแจกแจงแบบปกติของความคลาดเคลื่อนเช่นเดียวกับการวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย ซึ่งจะส่งผลให้เวกเตอร์ของค่าสังเกต $\mathbf{Y}$ มีการแจกแจงแบบปกติ โดยมีค่าเฉลี่ยเป็น $\mathbf{X}\beta$ และความแปรปรวนเป็น $\sigma^2\mathbf{I}$ เขียนโดยรวมได้ดังนี้ $\mathbf{Y} \sim NID\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$

3.3.1 ช่วงความเชื่อมั่นของ β_j (Confidence Interval of β_j)

เนื่องจากตัวประมาณกำลังสองน้อยที่สุด $\mathbf{b}$ เป็นฟังก์ชันเชิงเส้นของค่าสังเกต ดังนั้นจะมีการแจกแจงแบบปกติด้วยเช่นเดียวกัน โดยมีเวกเตอร์ค่าเฉลี่ย β และเมตริกซ์ความแปรปรวน $\sigma^2(\mathbf{X'X})^{-1}$ สามารถเขียนโดยรวมได้เป็น $\mathbf{b} \sim NID\left(\beta, \sigma^2(\mathbf{X'X})^{-1}\right)$ ซึ่งชี้ว่าค่าสัมประสิทธิ์ถดถอยแต่ละตัว b_j จะมีการแจกแจงแบบปกติด้วยค่าเฉลี่ย β_j และความแปรปรวน $\sigma^2 C_{jj}$ เมื่อ C_{jj} เป็นค่าที่อยู่บนเส้นทแยงมุมของเมตริกซ์ $(\mathbf{X'X})^{-1}$ แต่ในทางปฏิบัติมักไม่ทราบ σ^2 และจะประมาณด้วย S^2 ดังนั้นตัวสถิติ

$$
\frac{b_j - \beta_j}{S_{b_j}} \quad \sim \quad t_{n-k-1}, \qquad j = 0, 1, \ldots, k
$$

เมื่อ $S_{b_j} = \sqrt{S^2 \cdot C_{jj}}$ และเรียก S_{b_j} ว่าเป็นค่าความคลาดเคลื่อนมาตรฐานของการประมาณ β_j (Standard error of estimate for β_j)

ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ β_j คือ

$$
b_j \pm t_{\frac{\alpha}{2}, n-k-1} \cdot S_{b_j} \tag{3.18}
$$

ตัวอย่างที่ 3.3 จงสร้างช่วงความเชื่อมั่น 90% ของพารามิเตอร์ β_0, β_1 และ β_2 โดยใช้ข้อมูลในตัวอย่างที่ 3.1

วิธีทำ จากข้อมูลในตัวอย่างที่ 3.1 ได้

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix}$$

$$S^2 = MSE = 14.274$$

เปิดตารางค่าวิกฤตได้ดังนี้ $t_{0.025,12} = 1.782$

ดังนั้นช่วงความเชื่อมั่น 90% ของ β_0 คือ

$$\begin{aligned} b_0 \pm t_{\frac{\alpha}{2},n-k-1} \cdot \sqrt{S^2 \cdot C_{00}} &= 6.9652 \pm (1.782)\left(\sqrt{(14.274)(0.823589)}\right) \\ &= 6.9652 \pm 3.4287 \\ \text{หรือ} &\quad (3.5365,\ 10.3939) \end{aligned}$$

ดังนั้นช่วงความเชื่อมั่น 90% ของ β_1 คือ

$$\begin{aligned} b_1 \pm t_{\frac{\alpha}{2},n-k-1} \cdot \sqrt{S^2 \cdot C_{11}} &= 1.2974 \pm (1.782)\left(\sqrt{(14.274)(0.0031291)}\right) \\ &= 1.2974 \pm 0.3766 \\ \text{หรือ} &\quad (0.9208,\ 1.6740) \end{aligned}$$

ดังนั้นช่วงความเชื่อมั่น 90% ของ β_2 คือ

$$\begin{aligned} b_2 \pm t_{\frac{\alpha}{2},n-k-1} \cdot \sqrt{S^2 \cdot C_{22}} &= 3.2161 \pm (1.782)\left(\sqrt{(14.274)(0.0072613)}\right) \\ &= 3.2161 \pm 0.5737 \\ \text{หรือ} &\quad (2.6424,\ 3.7898) \end{aligned}$$

3.3.2 การทดสอบสมมติฐานเกี่ยวกับ β_j (Hypothesis Testing Concerning β_j)

การทดสอบสมมติฐานเกี่ยวกับค่าสัมประสิทธิ์ถดถอย β_j มีประโยชน์ในการกำหนดอิทธิพลของตัวแปรอิสระแต่ละตัวในโมเดล เช่น โมเดลอาจมีประสิทธิภาพมากขึ้นถ้าเพิ่มตัวแปรอิสระเข้าไปในโมเดล หรือตัดตัวแปรอิสระหนึ่งตัวหรือมากกว่าออกจากโมเดล มีขั้นตอนดังนี้

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0, \qquad j = 0, 1, \ldots, k$$

สถิติทดสอบ:

$$t_c = \frac{b_j}{S_{b_j}} = \frac{b_j}{\sqrt{S^2 \cdot C_{jj}}} \tag{3.19}$$

เมื่อ C_{jj} เป็นสมาชิกที่อยู่บนเส้นทแยงมุมของ $\mathbf{X'X})^{-1}$ ที่สอดคล้องกับ b_j

บริเวณวิกฤติ: ที่ระดับนัยสำคัญ α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|t_c| \geq t_{\alpha/2, n-k-1}$

ยอมรับ H_0 ถ้า $|t_c| < t_{\alpha/2, n-k-1}$

จะเห็นได้ว่าการทดสอบ $H_0 : \beta_j = 0$, $j = 0, 1, \ldots, k$ เป็นการทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวแบบบางส่วน (Partial test or marginal test) เนื่องจากค่าสัมประสิทธิ์ถดถอย b_j ขึ้นอยู่กับตัวแปรอิสระอื่นทั้งหมดที่อยู่ในโมเดล ดังนั้นจึงถือเป็นการทดสอบอิทธิพลของ X_j เมื่อกำหนดให้ตัวแปรอิสระอื่นอยู่ในโมเดล

ตัวอย่างที่ 3.4 จากข้อมูลในตัวอย่างที่ 3.1 ที่ระดับนัยสำคัญ 0.10

1. จงทดสอบว่าสมการถดถอยผ่านจุดกำเนิดเหมาะสมกับข้อมูลชุดนี้หรือไม่
2. จงทดสอบอิทธิพลของระดับสติปัญญาต่อความพึงพอใจในการทำงานของพนักงานในโรงพยาบาลแห่งนี้ เมื่อกำหนดให้ดัชนีการปรับตัวอยู่ในโมเดล
3. จงทดสอบอิทธิพลของดัชนีการปรับตัวต่อความพึงพอใจในการทำงานของพนักงาน เมื่อกำหนดให้ระดับสติปัญญาอยู่ในโมเดล

วิธีทำ จาก

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix}$$

$$S^2 = MSE = 14.274$$

$$t_{0.025,12} = 1.782$$

1. กำหนดสมมติฐานของการทดสอบเป็น

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

คำนวณสถิติทดสอบ:

$$t_c = \frac{b_0}{S_{b_0}} = \frac{6.9652}{\sqrt{(14.274)(0.823589)}} = \frac{6.9652}{3.4287} = 2.0314$$

บริเวณวิกฤติ: $t_{0.05,12} = 1.782$

สรุปผล: ปฏิเสธ H_0 นั่นคือ สมการถดถอยผ่านจุดกำเนิดไม่เหมาะสมกับข้อมูลชุดนี้ ที่ระดับ 0.10

2. กำหนดสมมติฐานของการทดสอบเป็น

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

คำนวณสถิติทดสอบ:

$$t_c = \frac{b_1}{S_{b_1}} = \frac{1.2974}{\sqrt{(14.274)(0.0031291)}} = \frac{1.2974}{0.2113} = 6.1389$$

บริเวณวิกฤติ: $t_{0.05,12} = 1.782$

สรุปผล: ปฏิเสธ H_0 นั่นคือ ระดับสติปัญญามีอิทธิพลต่อความพึงพอใจในการทำงานของพนักงานในโรงพยาบาลอย่างมีนัยสำคัญ ที่ระดับ 0.10 เมื่อกำหนดให้ดัชนีการปรับตัวอยู่ในโมเดล

3. กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_2 = 0 \quad vs. \quad H_1 : \beta_2 \neq 0$$

คำนวณสถิติทดสอบ:

$$t_c = \frac{b_2}{S_{b_2}} = \frac{3.2161}{\sqrt{(14.274)(0.0072613)}} = \frac{3.2161}{0.3219} = 9.9896$$

บริเวณวิกฤติ: $t_{0.05,12} = 1.782$

สรุปผล: ปฏิเสธ H_0 นั่นคือ ดัชนีการปรับตัวมีอิทธิพลต่อความพึงพอใจในการทำงานของพนักงานอย่างมีนัยสำคัญ ที่ระดับ 0.10 เมื่อกำหนดให้ระดับสติปัญญาอยู่ในโมเดล

3.4 การประมาณช่วงความเชื่อมั่นของ $E(\mathbf{Y} \mid \mathbf{X} = \mathbf{X}_0)$

ทำนองเดียวกับการวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย บางครั้งผู้วิจัยอาจสนใจที่จะประมาณค่าเฉลี่ยของการแจกแจงของ Y เมื่อกำหนดค่าของตัวแปรอิสระ $\mathbf{X}_0$ ให้ กำหนดเวกเตอร์ของ $\mathbf{X}_0$ เป็น

$$\mathbf{X}_0 = \begin{bmatrix} 1 \\ X_{01} \\ X_{02} \\ \vdots \\ X_{0k} \end{bmatrix}$$

เมื่อ $X_{01}, X_{02}, \ldots, X_{0k}$ แทน ค่าของตัวแปรอิสระตัวที่ $1, 2, \ldots, k$ ของค่าสังเกตค่าใหม่ตามลำดับ ค่าพยากรณ์ (Fitted value) ที่จุด $\mathbf{X} = \mathbf{X}_0$ คือ

$$\hat{\mathbf{Y}}_0 = \mathbf{X}_0'\mathbf{b} \tag{3.20}$$

ซึ่งเป็นตัวประมาณที่ไม่เอนเอียงของ $\mathbf{Y}_0$ เนื่องจาก

$$\begin{aligned} E(\hat{\mathbf{Y}}_0) &= E(\mathbf{X}_0'\mathbf{b}) \\ &= \mathbf{X}_0'E(\mathbf{b}) \\ &= \mathbf{X}_0'\beta \\ &= \mathbf{Y}_0 \end{aligned}$$

จากนิยามของความแปรปรวนของ $\hat{\mathbf{Y}}_0$ จะได้ว่า

$$\begin{aligned} V(\hat{\mathbf{Y}}_0) = \sigma^2_{\hat{\mathbf{Y}}_0} &= E\left(\hat{\mathbf{Y}}_0 - E(\mathbf{Y}_0 \mid \mathbf{X}_0)\right)\left(\hat{\mathbf{Y}}_0 - E(\mathbf{Y}_0 \mid \mathbf{X}_0)\right)' \\ &= E\left(\mathbf{X}_0'\mathbf{b} - \mathbf{X}_0'\beta\right)\left(\mathbf{X}_0'\mathbf{b} - \mathbf{X}_0'\beta\right)' \\ &= \mathbf{X}_0'E\left(\mathbf{b} - \beta\right)\left(\mathbf{b} - \beta\right)'\mathbf{X}_0 \\ &= \mathbf{X}_0'Cov(\mathbf{b})\mathbf{X}_0 \end{aligned}$$

ดังนั้นความแปรปรวนของการประมาณค่าเฉลี่ยของ $\mathbf{Y}$ เมื่อกำหนด $\mathbf{X} = \mathbf{X}_0$ คือ

$$V(\hat{\mathbf{Y}}_0) = \sigma^2_{\hat{\mathbf{Y}}_0} = \sigma^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$$

โดยทั่วไปไม่ทราบ σ^2 และต้องประมาณด้วย S^2 ได้ค่าประมาณของ $\sigma^2_{\hat{\mathbf{Y}}_0}$ คือ

$$S^2_{\hat{\mathbf{Y}}_0} = S^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \tag{3.21}$$

ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ $E(\mathbf{Y} \mid \mathbf{X} = \mathbf{X}_0)$ คือ

$$\hat{\mathbf{Y}}_0 \pm t_{\alpha/2,\, n-k-1} \cdot S_{\hat{\mathbf{Y}}_0} \tag{3.22}$$

3.5 การพยากรณ์ค่าสังเกตค่าใหม่ (Prediction of New Observation)

เป็นที่ทราบแล้วว่าโมเดลถดถอยสามารถใช้ในการพยากรณ์ค่าสังเกตค่าใหม่ที่สอดคล้องกับค่าของตัวแปรอิสระที่กำหนดให้ได้ โดยตัวประมาณแบบจุดของค่าสังเกตค่าใหม่ Y_0 ที่จุด $X_{01}, X_{02}, \ldots, X_{0k}$ คือ

$$\hat{\mathbf{Y}}_0 = \mathbf{X}_0'\mathbf{b}$$

และค่าความแปรปรวนของการพยากรณ์ค่า $\mathbf{Y}_0$ สามารถแสดงได้ดังนี้

$$\begin{aligned}
\sigma^2_{\mathbf{Y}_0} &= V(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0) \\
&= V(\mathbf{Y}_0) + V(\hat{\mathbf{Y}}_0) \\
&= \sigma^2 + \sigma^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \\
&= \sigma^2\left(1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right)
\end{aligned}$$

ประมาณ σ^2 โดยใช้ S^2 จะได้

$$S^2_{\mathbf{Y}_0} = S^2\left(1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right) \tag{3.23}$$

ดังนั้นช่วงแห่งการพยากรณ์ $(1-\alpha)100\%$ ของค่าสังเกต Y_0 ที่จุด $\mathbf{X} = \mathbf{X}_0$ คือ

$$\hat{\mathbf{Y}}_0 \pm t_{\alpha/2,\, n-k-1} \cdot S_{\mathbf{Y}_0} \tag{3.24}$$

ช่วงแห่งการพยากรณ์ใน (3.24) เป็นกรณีทั่วไปของช่วงแห่งการพยากรณ์ค่าสังเกตค่าใหม่ในโมเดลถดถอยเชิงเส้นตรงอย่างง่าย

ตัวอย่างที่ 3.5 จากข้อมูลในตัวอย่างที่ 3.1

1. จงสร้างช่วงความเชื่อมั่น 90% ของคะแนนความพึงพอใจเฉลี่ยสำหรับพนักงานที่มีคะแนนระดับสติปัญญาเป็น 13 และดัชนีการปรับตัวเป็น 7
2. จงสร้างช่วงแห่งการพยากรณ์ 90% ของคะแนนความพึงพอใจสำหรับพนักงานที่มีคะแนนระดับสติปัญญาเป็น 13 และดัชนีการปรับตัวเป็น 7

วิธีทำ จาก

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix}$$

$$S^2 = MSE = 14.274$$

$$t_{0.025,12} = 1.782$$

1. ในที่นี้ $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 13 \\ 7 \end{bmatrix}$

   ค่าพยากรณ์ของคะแนนความพึงพอใจสำหรับพนักงานที่มีคะแนนระดับสติปัญญาเป็น 13 และดัชนีการปรับตัวเป็น 7 คือ

$$\hat{\mathbf{Y}}_0 = \mathbf{X}_0'\mathbf{b} = \begin{bmatrix} 1 & 13 & 7 \end{bmatrix} \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix} = 46.3441$$

ค่าประมาณของความแปรปรวนของ $\hat{\mathbf{Y}}_0$ คือ

$$\begin{aligned}
S^2_{\hat{\mathbf{Y}}_0} &= S^2\,\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \\
&= 14.274 \begin{bmatrix} 1 & 13 & 7 \end{bmatrix} \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix} \begin{bmatrix} 1 \\ 13 \\ 7 \end{bmatrix} \\
&= (14.274)(0.0943) \\
&= 1.3457
\end{aligned}$$

ดังนั้นช่วงความเชื่อมั่น 90% ของ $E(\mathbf{Y} \mid \mathbf{X} = \mathbf{X}_0)$ คือ

$$\begin{aligned}
\hat{\mathbf{Y}}_0 \pm t_{\alpha/2,n-k-1} \cdot S_{\hat{\mathbf{Y}}_0} &= 46.3441 \pm (1.782)(\sqrt{1.3457}) \\
&= 46.3441 \pm 2.0672 \\
\text{หรือ}\quad & (44.2769,\ 48.4113)
\end{aligned}$$

นั่นคือ ช่วงดังกล่าวจะคลุมคะแนนความพึงพอใจเฉลี่ยที่แท้จริงด้วยความเชื่อมั่น 90%

2. คำนวณค่าประมาณของความแปรปรวนของ $\mathbf{Y}_0$ คือ

$$\begin{aligned}
S^2_{\mathbf{Y}_0} &= S^2\,\left(1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right) \\
&= 14.274 \left(1 + \begin{bmatrix} 1 & 13 & 7 \end{bmatrix} \begin{bmatrix} 0.823589 & -0.0423122 & -0.0424984 \\ -0.042312 & 0.0031291 & 0.0004459 \\ -0.042498 & 0.0004459 & 0.0072613 \end{bmatrix} \begin{bmatrix} 1 \\ 13 \\ 7 \end{bmatrix}\right) \\
&= (14.274)(1.0943) \\
&= 15.6200
\end{aligned}$$

ดังนั้นช่วงแห่งการพยากรณ์ 90% ของ Y_0 ที่จุด $\mathbf{X} = \mathbf{X}_0$ คือ

$$\begin{aligned}
\mathbf{Y}_0 \pm t_{\alpha/2,n-k-1} \cdot S_{\mathbf{Y}_0} &= 46.3441 \pm (1.782)(\sqrt{15.6200}) \\
&= 46.3441 \pm 7.0429 \\
\text{หรือ}\quad & (39.3012,\ 53.3870)
\end{aligned}$$

นั่นคือ ช่วงดังกล่าวจะคลุมคะแนนความพึงพอใจที่แท้จริงด้วยความเชื่อมั่น 90%

3.6 การทดสอบนัยสำคัญของสมการถดถอย (Test for Significance of Regression)

การทดสอบนัยสำคัญของสมการถดถอยเป็นการตรวจสอบว่าตัวแปรตาม Y มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระอย่างน้อยหนึ่งตัวหรือไม่ ซึ่งต้องอาศัยข้อกำหนดเบื้องต้นเกี่ยวกับการแจกแจงแบบปกติของความคลาดเคลื่อน เช่นเดียวกับการทดสอบสมมติฐานของค่าสัมประสิทธิ์ถดถอยแต่ละตัว ในที่นี้สามารถกำหนดสมมติฐานของการทดสอบได้เป็น

$$
\begin{aligned}
H_0 &: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\
H_1 &: \quad \beta_j \neq 0 \quad \text{อย่างน้อย 1 ค่า}, \qquad j = 1, 2, \ldots, k
\end{aligned}
\tag{3.25}
$$

การทดสอบนี้เป็นกรณีทั่วไปของการวิเคราะห์ความแปรปรวนสำหรับการวิเคราะห์การถดถอยอย่างง่าย โดยแบ่งความผันแปรทั้งหมดของ Y ออกเป็น 2 ส่วน คือ ส่วนที่สามารถอธิบายได้ด้วยสมการถดถอย และส่วนที่ไม่สามารถอธิบายได้ด้วยสมการถดถอย แสดงในรูปของผลรวมกำลังสองได้ดังนี้

$$SST = SSR + SSE$$

หาก H_0 เป็นจริงแล้ว จะได้ว่า

$$\frac{SSR}{\sigma^2} \sim \chi^2_k$$

ซึ่งจำนวนองศาความเป็นอิสระของ χ^2 มีค่าเท่ากับจำนวนตัวแปรอิสระนั่นเอง ทำนองเดียวกันจะได้ว่า

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-k-1}$$

โดย SSR และ SSE เป็นอิสระกัน ดังนั้นการทดสอบ (3.25) ทำได้โดยคำนวณสถิติทดสอบ

$$F_c = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} \tag{3.26}$$

ปฏิเสธ H_0 ถ้า $F_c > F_{k,n-k-1(\alpha)}$ ซึ่งแสดงว่ามีตัวแปรอิสระอย่างน้อยหนึ่งตัวที่มีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ขั้นตอนการทดสอบสามารถสรุปในตาราง ANOVA ได้ดังนี้

ANOVA

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

เมื่อ

$$SST = S_{yy} = \mathbf{Y}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \tag{3.27}$$

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \tag{3.28}$$

$$SSE = SST - SSR \tag{3.29}$$

ตัวอย่างที่ 3.6 ที่ระดับนัยสำคัญ 0.10 จงทดสอบนัยสำคัญของสมการถดถอยในตัวอย่างที่ 3.1

วิธีทำ จาก

$$\mathbf{b} = \begin{bmatrix} 6.9652 \\ 1.2974 \\ 3.2161 \end{bmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 598 \\ 8,009 \\ 3,451 \end{bmatrix},$$

$$\sum_{i=1}^{n} Y_i = 598, \qquad \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2 = 25,826, \qquad n = 15$$

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0: \ \beta_1 = \beta_2 = 0$$

$$H_1: \ \beta_j \neq 0 \ \text{ อย่างน้อย 1 ค่า}, \qquad j = 1, 2$$

คำนวณค่าต่าง ๆ ได้ดังนี้

$$SST = \mathbf{Y}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 25,826 - \frac{(598)^2}{15} = 1,985.7333$$

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 25,654.7 - \frac{(598)^2}{15} = 1,814.4333$$

$$SSE = SST - SSR = 1,985.7333 - 1,814.4333 = 171.2999$$

แทนค่าลงในตารางวิเคราะห์ความแปรปรวน

ANOVA

Source of variation	*df*	*SS*	*MS*	*F*
Regression	2	1,814.4333	907.2167	$F_c = 63.5529$
Error	12	171.2999	14.2750	
Total	14	1,985.7333		

เนื่องจาก $F_c = 63.5529 > F_{2,12(0.10)} = 2.81$ จึงปฏิเสธ H_0 นั่นคือ ความพึงพอใจมีความสัมพันธ์กับระดับสติปัญญา และ/หรือ ดัชนีการปรับตัว อย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.10 แต่อย่างไรก็ตามไม่ได้หมายความว่ารูปแบบความสัมพันธ์ที่ได้นั้นเหมาะสม จะต้องมีการทดสอบความเหมาะสมของโมเดลต่อไป

3.7 การทดสอบสมมติฐานโดยใช้ Extra-sum-of-squares

การทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวที่มีต่อตัวแปรตาม Y เมื่อให้ตัวแปรอิสระอื่นอยู่ในสมการ นอกจากจะใช้สถิติ t ในการทดสอบแล้ว ยังสามารถใช้วิธี Extra-sum-of-squares ในการทดสอบดังกล่าวได้อีกด้วย ซึ่งวิธี Extra-sum-of-squares ยังสามารถใช้ในการตรวจสอบอิทธิพลของสับเซ็ตของตัวแปรอิสระที่มีต่อตัวแปรตาม Y ได้เช่นกัน พิจารณาโมเดลถดถอยที่มีตัวแปรอิสระ k ตัว

$$\underset{(n\times1)}{\mathbf{Y}} = \underset{(n\times p)}{\mathbf{X}}\underset{(p\times1)}{\beta} + \underset{(n\times1)}{\epsilon}, \qquad p = k+1$$

ในที่นี้สนใจทดสอบอิทธิพลของสับเซ็ตของตัวแปรอิสระ r ตัว $(r < k)$ ว่ามีอิทธิพลต่อโมเดลถดถอยอย่างมีนัยสำคัญหรือไม่ การทดสอบทำได้โดยแบ่งเมตริกซ์ของตัวแปรอิสระและเวกเตอร์ของค่าสัมประสิทธิ์ถดถอยออกเป็น 2 ส่วน ดังนี้

$$\mathbf{X} = \left[\underset{n\times(p-r)}{\mathbf{X}_1} \;\middle|\; \underset{n\times r}{\mathbf{X}_2}\right], \qquad \beta = \begin{bmatrix} \underset{(p-r)\times 1}{\beta_1} \\ \hline \underset{r\times 1}{\beta_2} \end{bmatrix}$$

เมื่อ $\mathbf{X}_1$, และ $\mathbf{X}_2$ แทน คอลัมน์ของ $\mathbf{X}$ ที่สอดคล้องกับ β_1 และ β_2 ตามลำดับ ดังนั้นสามารถเขียนโมเดลถดถอยอีกรูปแบบหนึ่งได้เป็น

$$\mathbf{Y} \;=\; \mathbf{X}\beta + \epsilon \;=\; \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \tag{3.30}$$

โดยจะเรียกสมการ (3.30) ว่า *Full model* เป็นที่ทราบแล้วว่าตัวประมาณกำลังสองน้อยที่สุดของค่าสัมประสิทธิ์ถดถอยใน Full model คือ

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

และได้

$$\begin{aligned}
SST &= \mathbf{Y}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} && (df = n-1) \\
SSR(\mathbf{X}_1, \mathbf{X}_2) &= \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} && (df = k) \\
SSE(\mathbf{X}_1, \mathbf{X}_2) &= \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} && (df = n-k-1)
\end{aligned}$$

เมื่อ

SST แทน ผลรวมกำลังสองทั้งหมดใน Full model

$SSR(\mathbf{X}_1, \mathbf{X}_2)$ แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอยใน Full model

$SSE(\mathbf{X}_1, \mathbf{X}_2)$ แทน ผลรวมกำลังสองเนื่องมาจากความคลาดเคลื่อนใน Full model

หากสมมติฐานที่ต้องการทดสอบ คือ

$$H_0 : \beta_2 = \mathbf{0} \qquad vs. \qquad H_1 : \beta_2 \neq \mathbf{0}$$

ถ้า H_0 เป็นจริง สมการ (3.30) จะลดรูปเป็น

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \epsilon \tag{3.31}$$

เรียกสมการ (3.31) ว่า *Reduced model* และได้ตัวประมาณกำลังสองน้อยที่สุดของ β_1 คือ

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} \tag{3.32}$$

ค่าผลรวมกำลังสองเนื่องมาจากสมการถดถอยและความคลาดเคลื่อนใน Reduced model คือ

$$\begin{aligned} SSR(\mathbf{X}_1) &= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} && (df = k-r) \\ SSE(\mathbf{X}_1) &= SST(\mathbf{X}_1, \mathbf{X}_2) - SSR(\mathbf{X}_1) && (df = n-(k-r)-1) \end{aligned}$$

เมื่อ

$SSR(\mathbf{X}_1)$ แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอยใน Reduced model

$SSE(\mathbf{X}_1)$ แทน ผลรวมกำลังสองเนื่องมาจากความคลาดเคลื่อนใน Reduced model

การทดสอบสมมติฐาน $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ ซึ่งเป็นการทดสอบอิทธิพลของ $\mathbf{X}_2$ เมื่อกำหนดให้ $\mathbf{X}_1$ อยู่ในโมเดล ทำได้โดยคำนวณค่า

$$SSR(\mathbf{X}_2 \mid \mathbf{X}_1) = SSR(\mathbf{X}_1, \mathbf{X}_2) - SSR(\mathbf{X}_1) \tag{3.33}$$

หรือ

$$SSR(\mathbf{X}_2 \mid \mathbf{X}_1) = SSR_{Full} - SSR_{Reduced} \tag{3.34}$$

ซึ่งมีจำนวนองศาความเป็นอิสระเป็น $k-(k-r)=r$ โดยจะเรียกผลรวมกำลังสองใน (3.33) ว่า *Extra-sum-of-squares* เนื่องจาก $\mathbf{X}_2$ และเป็นค่าที่ใช้วัดผลรวมกำลังสองของสมการถดถอยที่เพิ่มขึ้นเนื่องจากการเพิ่มตัวแปรอิสระ $X_{k-r+1}, X_{k-r+2}, \ldots, X_k$ เข้าไปในโมเดลที่มี $X_1, X_2, \ldots, X_{k-r}$ อยู่ นอกจากนี้พบว่าค่า $SSR(\mathbf{X}_2 \mid \mathbf{X}_1)$ เป็นอิสระจาก MSE ใน Full model ดังนั้นการทดสอบสมมติฐาน $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ ทำได้โดยคำนวณสถิติทดสอบ

$$F_c = \frac{SSR(\mathbf{X}_1 \mid \mathbf{X}_2)\Big/r}{SSE(\mathbf{X}_1, \mathbf{X}_2)\Big/(n-k-1)} \tag{3.35}$$

เมื่อ r แทน ผลต่างของจำนวนองศาความเป็นอิสระระหว่าง Full model และ Reduced model

เรียกการทดสอบใน (3.35) ว่า *Partial F-test* โดยจะปฏิเสธ H_0 ถ้า $F_c > F_{r,n-k-1(\alpha)}$ ซึ่งแสดงว่ามีพารามิเตอร์

อย่างน้อยหนึ่งตัวใน β_2 ที่ไม่เท่ากับศูนย์ นั่นคือ มีตัวแปรอิสระ $X_{k-r+1}, X_{k-r+2}, \ldots, X_k$ อย่างน้อยหนึ่งตัวใน $\mathbf{X}_2$ ที่มีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ที่ระดับ 0.10

นอกจากนี้การคำนวณค่า $SSR(\mathbf{X}_2 \mid \mathbf{X}_1)$ ในสมการ (3.33) สามารถทำได้โดยใช้ SSE เนื่องจาก

$$\begin{aligned} SSR(\mathbf{X}_1, \mathbf{X}_2) &= SST - SSE(\mathbf{X}_1, \mathbf{X}_2) \\ SSR(X_1) &= SST - SSE(\mathbf{X}_1) \end{aligned}$$

และได้ว่า

$$\begin{aligned} SSR(\mathbf{X}_2 \mid \mathbf{X}_1) &= SSR(\mathbf{X}_1, \mathbf{X}_2) - SSR(\mathbf{X}_1) \\ &= (SST - SSE(\mathbf{X}_1, \mathbf{X}_2)) - (SST - SSE(\mathbf{X}_1)) \\ &= SSE(\mathbf{X}_1) - SSE(\mathbf{X}_1, \mathbf{X}_2) \end{aligned} \tag{3.36}$$

หรือ

$$SSR(\mathbf{X}_2 \mid \mathbf{X}_1) = SSE_{Reduced} - SSE_{Full} \tag{3.37}$$

ดังนั้นสมการ (3.35) สามารถเขียนใหม่ได้เป็น

$$F_c = \frac{(SSR_{Full} - SSR_{Reduced}) \big/ r}{MSE_{Full}} = \frac{(SSE_{Reduced} - SSE_{Full}) \big/ r}{MSE_{Full}} \tag{3.38}$$

โดยทั่วไปสามารถคำนวณค่า $SSR(X_j \mid X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k), \quad 1 \leq j \leq k$ เพื่อใช้วัดการเพิ่มขึ้นของ SSR เนื่องจากการเพิ่ม X_j เข้าไปในโมเดลที่มี $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ อยู่ในโมเดลแล้ว ซึ่งการทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์แต่ละตัวด้วย Partial F-test จะเทียบเท่ากับการทดสอบโดยใช้สถิติทดสอบทีใน (3.19) แต่อย่างไรก็ตามการทดสอบด้วย Partial F-test จะมีลักษณะทั่วไปมากกว่า โดยสามารถใช้วัดอิทธิพลของกลุ่มของตัวแปรได้อีกด้วย นอกจากนี้การทดสอบด้วย Partial F-test มีบทบาทที่สำคัญ โดยเฉพาะอย่างยิ่งในการสร้างโมเดล (Model building) เพื่อค้นหากลุ่มของตัวแปรอิสระที่ดีที่สุดในโมเดลนั้น ๆ จึงทำให้วิธี Extra-sum-of-squares ค่อนข้างเป็นที่นิยม

บางกรณีพบว่าตัวแปรอิสระมีลักษณะเรียงลำดับตามธรรมชาติ พิจารณาโมเดลต่อไปนี้

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

ผู้วิจัยอาจสนใจทดสอบอิทธิพลเชิงเส้น (Linear effect) ของ X โดยคำนวณค่า $SSR(X)$ และอิทธิพลของเทอมกำลังสอง (Quadratic effect) ของ X โดยคำนวณค่า $SSR(X^2 \mid X)$ เพื่อดูว่าการเพิ่มเทอมกำลังสองเข้าไปในโมเดลที่มีเทอมกำลังหนึ่งอยู่แล้วมีส่วนช่วยในการพยากรณ์หรือไม่ ดังนั้นการเพิ่มตัวแปรเข้าไปในโมเดลทีละตัวแล้วตรวจสอบการมีส่วนช่วยในการพยากรณ์ของตัวแปรที่เพิ่มเข้าไปทีละขั้น เมื่อให้ตัวแปรอิสระอื่นที่เพิ่มเข้าไปก่อนหน้าอยู่ในโมเดลทั้งหมด สามารถทำได้ด้วยวิธี Extra-sum-of-squares โดยแบ่ง SSR ออกเป็นส่วนประกอบเดี่ยว ๆ ที่มีจำนวนองศาความเป็นอิสระเป็น 1 เพื่อทดสอบอิทธิพลของตัวแปรอิสระที่เพิ่มเข้าไปแต่ละตัว

พิจารณาโมเดล

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

จากหลักการวิเคราะห์ความแปรปรวน จะได้ว่า

$$SST = SSR(X_1, X_2, X_3) + \epsilon$$

โดย $SSR(X_1, X_2, X_3)$ สามารถแบ่งเป็นส่วนย่อย ๆ ได้ 3 ส่วน แต่ละส่วนมี $df = 1$ ดังนี้

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2)$$

ซึ่งลำดับของตัวแปรอิสระในแต่ละส่วนประกอบย่อยทางด้านขวามือของสมการข้างต้นไม่เจาะจง ดังนั้นอาจเขียน $SSR(X_1, X_2, X_3)$ ใหม่ได้เป็น

$$SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_1 \mid X_2) + SSR(X_3 \mid X_1, X_2)$$

หรือ

$$SSR(X_1, X_2, X_3) = SSR(X_3) + SSR(X_2 \mid X_3) + SSR(X_1 \mid X_2, X_3)$$

อย่างไรก็ตามวิธี Extra-sum-of-squares ไม่สามารถใช้แบ่งผลรวมกำลังสองเนื่องจากสมการถดถอยได้เสมอ โดยทั่วไปแล้วพบว่า

$$SSR(X_1, X_2, X_3) \neq SSR(X_1 \mid X_2, X_3) + SSR(X_2 \mid X_1, X_3) + SSR(X_3 \mid X_1, X_2)$$

ตัวอย่างที่ 3.7 จากข้อมูลในตัวอย่างที่ 3.1 ที่ระดับนัยสำคัญ 0.10 จงตรวจสอบอิทธิพลของดัชนีการปรับตัวที่มีต่อความพึงพอใจในการทำงาน สำหรับโมเดลที่มีระดับสติปัญญาอยู่แล้ว โดยใช้ Partial F-test

วิธีทำ กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_2 = 0 \quad vs. \quad H_1 : \beta_2 \neq 0$$

เพื่อที่จะทดสอบสมมติฐานข้างต้น ต้องคำนวณค่า

$$SSR(X_2 \mid X_1) = SSR(X_1, X_2) - SSR(X_1)$$

จากตัวอย่างที่แล้วทราบว่า

$$\begin{aligned} SSR(X_1, X_2) &= 1,814.4333 \qquad (df = 2) \\ SSE(X_1, X_2) &= 171.2999 \qquad (df = 12) \end{aligned}$$

หาก H_0 เป็นจริง จะได้ Reduced model เป็น $Y = \beta_0 + \beta_1 X_1 + \epsilon$ โดย

$$\begin{aligned} \mathbf{X}_1'\mathbf{X}_1 &= \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 \end{bmatrix} = \begin{bmatrix} 15 & 192 \\ 192 & 2,780 \end{bmatrix} \\ \mathbf{X}_1'\mathbf{Y} &= \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1} Y_i \end{bmatrix} = \begin{bmatrix} 598 \\ 8,009 \end{bmatrix} \end{aligned}$$

ประมาณค่า β_0 และ β_1 ใน Reduced model ได้สมการดังนี้

$$\hat{Y} = 25.788 + 1.0999 X_1$$

คำนวณค่าผลรวมกำลังสองเนื่องมาจากสมการถดถอยใน Reduced model ได้เป็น

$$
\begin{aligned}
SSR(X_1) &= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \\
&= \begin{bmatrix} 25.788 & 1.0999 \end{bmatrix} \begin{bmatrix} 598 \\ 8,009 \end{bmatrix} - \frac{(598)^2}{15} \\
&= 24,230.3231 - 23,840.2667 \\
&= 390.0564 \qquad\qquad (df = 1)
\end{aligned}
$$

ดังนั้นจะได้ว่า

$$SSR(X_2 \mid X_1) = 1,814.4333 - 390.0564 = 1,424.3769$$

คำนวณ Partial F-test ได้ดังนี้

$$
\begin{aligned}
F_c &= \frac{SSR(X_2 \mid X_1)/r}{SSE(X_1, X_2)/(n-k-1)} \\
&= \frac{1,424.3769/1}{171.2999/12} \\
&= 99.7813
\end{aligned}
$$

แทนค่าลงในตารางวิเคราะห์ความแปรปรวน

ANOVA

Source of variation	df	SS	MS	F
Regression	2	1,814.4333	907.2167	63.5529
X_1	1	390.0564	390.0564	27.3244
$X_2 \mid X_1$	1	1,424.3769	1,424.3769	99.7813
Error	12	171.2999	14.2750	
Total	14	1,985.7333		

เนื่องจาก $F_c = 99.7813 > F_{1,12(0.10)} = 3.18$ จึงปฏิเสธ H_0 นั่นคือ ความพึงพอใจมีความสัมพันธ์กับดัชนีการปรับตัวอย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.10 เมื่อให้ระดับสติปัญญามีค่าคงที่

จะเห็นได้ว่าการทดสอบโดยใช้ Partial F-test ในตัวอย่างที่ 3.7 เป็นการทดสอบอิทธิพลของตัวแปรอิสระเพียงตัวเดียว ซึ่งจะให้ผลการทดสอบเหมือนกับการใช้สถิติทดสอบที จากตัวอย่างที่ 3.4 การทดสอบสมมติฐาน $H_0 : \beta_2 = 0$ โดยใช้สถิติทดสอบที พบว่า $t_c = 9.9896$ และเนื่องจาก $t_\nu^2 = F_{1,\nu}$ ดังนั้น

$$t_c^2 = (9.9896)^2 = 99.7921 \simeq F_c$$

3.8 ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุ (Coefficient of Multiple Determination)

ค่าสัมประสิทธิ์ตัวกำหนดแบบพหุแทนด้วย R^2 ซึ่งถูกกำหนดโดย

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{3.39}$$

R^2 เป็นค่าที่ใช้วัดความผันแปรทั้งหมดของ Y ที่สามารถอธิบายได้ด้วยตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ นอกจากนี้ยังสามารถแสดงได้ว่า R^2 เกิดจากกำลังสองของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างเวกเตอร์ของค่าสังเกต $\mathbf{Y}$ และเวกเตอร์ของค่าพยากรณ์ $\hat{\mathbf{Y}}$ ทำนองเดียวกับการวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย นั่นคือ $0 \leq R^2 \leq 1$ แต่อย่างไรก็ตามค่า R^2 ที่สูง ไม่ได้หมายความว่าสมการถดถอยเชิงเส้นตรงนั้นเป็นสมการที่ดีเสมอไป เนื่องจากการเพิ่มตัวแปรอิสระเข้าไปในโมเดล จะทำให้ R^2 มีค่าเพิ่มขึ้นเสมอ ไม่ว่าตัวแปรอิสระที่เพิ่มเข้าไปนั้นมีส่วนช่วยในการอธิบายตัวแปรตาม Y หรือไม่ก็ตาม ดังนั้นจึงเป็นไปได้ที่โมเดลที่มีค่า R^2 สูง ๆ อาจให้ค่าประมาณหรือค่าพยากรณ์ที่แตกต่างจากค่าจริงมาก

นอกจากนี้รากที่สองที่เป็นบวกของ R^2 ก็คือ ค่าสัมประสิทธิ์สหสัมพันธ์เชิงพหุ (Multiple correlation coefficient) ระหว่าง Y กับ $X_1, X_2, \ldots, X_k$ นั่นเอง ซึ่งแทนด้วย R และเป็นค่าที่ใช้วัดความสัมพันธ์เชิงเส้นตรงระหว่าง Y กับ $X_1, X_2, \ldots, X_k$

เนื่องจาก R^2 มีค่าสูงขึ้นเสมอ เมื่อเพิ่มตัวแปรอิสระตัวใหม่เข้าไปในโมเดล ด้วยเหตุนี้บางครั้งจะพิจารณาค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่มีการปรับค่าแล้ว (Adjusted coefficient of multiple determination) แทนด้วย R^2_{adj} ซึ่งนิยมใช้สำหรับการเลือกตัวแปรอิสระในขั้นตอนของการสร้างโมเดล และช่วยป้องกันการเพิ่มเทอมที่ไม่จำเป็นเข้าในโมเดล (Overfitting) โดยมีสูตรที่ใช้ในการคำนวณดังนี้

$$\begin{aligned} R^2_{adj} &= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} && (3.40) \\ &= 1 - \frac{(n-1)}{(n-k-1)} \cdot \frac{SSE}{SST} && (3.41) \\ &= 1 - \frac{(n-1)}{(n-k-1)} \cdot (1 - R^2) && (3.42) \end{aligned}$$

ตัวอย่างที่ 3.8 จากข้อมูลในตัวอย่างที่ 3.1 จงหาค่าสัมประสิทธิ์ตัวกำหนดแบบพหุและค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่มีการปรับค่าแล้วของสมการถดถอยที่ได้ในตัวอย่างที่ 3.1

วิธีทำ จาก

$$SSR = 1,814.4333, \quad SST = 1,985.7333, \quad SSE = 171.2999, \quad n = 15, \quad k = 2$$

คำนวณค่าสัมประสิทธิ์ตัวกำหนดแบบพหุได้ดังนี้

$$R^2 = \frac{SSR}{SST} = \frac{1,814.4333}{1,985.7333} = 0.9137$$

และค่าสัมประสิทธิ์ตัวกำหนด แบบพหุที่มีการปรับค่าแล้ว

$$\begin{aligned} R^2_{adj} &= 1 - \frac{(n-1)}{(n-k-1)} \cdot \frac{SSE}{SST} \\ &= 1 - \frac{(15-1)}{(15-2-1)} \cdot \frac{171.2999}{1,985.7333} \\ &= 0.8993 \end{aligned}$$

จะเห็นได้ว่าค่า R^2 และ R^2_{adj} ของตัวอย่างที่ 3.8 มีค่าใกล้เคียงกัน แต่ถ้าค่า R^2 และ R^2_{adj} มีค่าแตกต่าง กันมาก อาจชี้ว่าสมการถดถอยได้รวมเอาเทอมที่ไม่มีส่วนช่วยในการอธิบายความสัมพันธ์เข้าไว้ด้วย

3.9 ค่าสัมประสิทธิ์ตัวกำหนดบางส่วน (Coefficient of Partial Determination)

Extra-sum-of-squares นอกจากใช้ในการทดสอบสมมติฐานเกี่ยวกับค่าสัมประสิทธิ์ถดถอยแล้ว ยังสามารถใช้วัดความสัมพันธ์ระหว่างตัวแปรได้อีกด้วย โดยใช้ในการคำนวณ *ค่าสัมประสิทธิ์ตัวกำหนดบางส่วน* เพื่ออธิบายอิทธิพลของตัวแปรอิสระตัวใดตัวหนึ่งต่อสมการถดถอย เมื่อตัวแปรอิสระอื่นทั้งหมดอยู่ในโมเดลด้วย ซึ่งแตกต่างจากค่าสัมประสิทธิ์ตัวกำหนดแบบพหุที่ใช้อธิบายความสัมพันธ์ระหว่างตัวแปรตาม Y กับเซ็ตของตัวแปรอิสระภายใต้การพิจารณาทุกตัว

พิจารณาโมเดลถดถอยกรณีที่มีตัวแปรอิสระ 2 ตัว

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

โดยค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่าง Y กับ X_1 เมื่อให้ X_2 อยู่ในโมเดล สามารถแสดงได้ด้วย

$r^2_{Y1.2}$ มีสูตรเป็น

$$r^2_{Y1.2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)}$$
$$= \frac{SSR(X_1 \mid X_2)}{SSE(X_2)} \tag{3.43}$$

เมื่อ

$SSE(X_2)$ แทน ความผันแปรของ Y เมื่อมี X_2 อยู่ในโมเดล

$SSE(X_1, X_2)$ แทน ความผันแปรของ Y เมื่อมี X_1, X_2 อยู่ในโมเดล

ซึ่ง $r^2_{Y1.2}$ เป็นค่าที่ใช้วัดความผันแปรของ Y ที่ลดลงเนื่องมาจาก X_1 เมื่อมี X_2 อยู่ในโมเดล โดยค่าสัมประสิทธิ์ตัวกำหนดบางส่วนมีค่าอยู่ระหว่าง 0 ถึง 1 เช่นเดียวกัน และการแปลความหมายของ $r^2_{Y1.2}$ จะเหมือนกับค่าสัมประสิทธิ์ตัวกำหนด R^2 ที่ใช้อธิบายความสัมพันธ์ระหว่างความคลาดเคลื่อนที่เกิดจากสมการถดถอยของ Y กับ X_2 และความคลาดเคลื่อนที่เกิดจากสมการถดถอยระหว่าง X_1 กับ X_2 หรือกล่าวอีกนัยหนึ่งได้ว่า $r^2_{Y1.2}$ เป็นค่าที่ใช้วัดความสัมพันธ์ระหว่าง Y กับ X_1 หลังจากที่ถูกปรับด้วยความสัมพันธ์เชิงเส้นกับ X_2 แล้วนั่นเอง ทำนองเดียวกันค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่าง Y กับ X_2 เมื่อให้ X_1 อยู่ในโมเดล สามารถเขียนได้เป็น

$$r^2_{Y2.1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$
$$= \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} \tag{3.44}$$

กรณีที่มีตัวแปรอิสระมากกว่า 2 ตัว การหาค่าสัมประสิทธิ์ตัวกำหนดบางส่วนสามารถทำได้ในลักษณะคล้ายคลึงกับกรณีที่มีตัวแปรอิสระ 2 ตัว เช่น พิจารณาโมเดลถดถอยที่มีตัวแปรอิสระ 3 ตัว คำนวณค่าสัมประสิทธิ์ตัวกำหนดบางส่วนได้เป็น

$$r^2_{Y1.23} = \frac{SSR(X_1 \mid X_2, X_3)}{SSE(X_2, X_3)}$$
$$r^2_{Y2.13} = \frac{SSR(X_2 \mid X_1, X_3)}{SSE(X_1, X_3)}$$
$$r^2_{Y3.12} = \frac{SSR(X_3 \mid X_1, X_2)}{SSE(X_1, X_2)}$$

ทำนองเดียวกันถ้ามีตัวแปรอิสระ 4 ตัว จะได้ว่า

$$\begin{aligned}
r^2_{Y1.234} &= \frac{SSR(X_1 \mid X_2, X_3, X_4)}{SSE(X_2, X_3, X_4)} \\
r^2_{Y2.134} &= \frac{SSR(X_2 \mid X_1, X_3, X_4)}{SSE(X_1, X_3, X_4)} \\
r^2_{Y3.124} &= \frac{SSR(X_3 \mid X_1, X_2, X_4)}{SSE(X_1, X_2, X_4)} \\
r^2_{Y4.123} &= \frac{SSR(X_4 \mid X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}
\end{aligned}$$

รากที่สองของค่าสัมประสิทธิ์ตัวกำหนดจะถูกเรียกว่า *ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน* (Coefficient of partial correlation) โดยจะมีเครื่องหมายเหมือนกับค่าสัมประสิทธิ์ถดถอยที่สอดคล้องกัน และนิยมใช้สำหรับการเลือกตัวแปรอิสระในขั้นตอนของการสร้างโมเดล

ตัวอย่างที่ 3.9 จากข้อมูลในตัวอย่างที่ 3.1 จงหาค่าสัมประสิทธิ์ตัวกำหนดบางส่วนและค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน

วิธีทำ จาก

$$SST = 1,985.7333, \quad SSR(X_1, X_2) = 1,814.4333, \quad SSE(X_1, X_2) = 171.2999,$$
$$SSR(X_1) = 390.0564, \qquad SSE(X_1) = 1,595.6769,$$
$$SSR(X_2 \mid X_1) = 1,424.3769, \qquad n = 15$$

สร้างสมการถดถอยระหว่าง Y กับ X_2 ได้เป็น

$$\hat{Y} = 24.509 + 3.0312X_2$$

คำนวณค่าผลรวมกำลังสองเนื่องมาจากสมการถดถอยระหว่าง Y กับ X_2 ได้เป็น

$$
\begin{aligned}
SSR(X_2) &= \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \\
&= \begin{bmatrix} 24.509 & 3.0312 \end{bmatrix} \begin{bmatrix} 598 \\ 3491 \end{bmatrix} - \frac{(598)^2}{15} \\
&= 25,238.3012 - 23,840.2667 \\
&= 1,398.0345 \qquad (df = 1) \\
SSE(X_2) &= SST - SSR(X_2) \\
&= 1,985.7333 - 1,398.0345 \\
&= 587.6988 \qquad (df = 13)
\end{aligned}
$$

ดังนั้นจะได้ว่า

$$
\begin{aligned}
SSR(X_1 \mid X_2) &= SSR(X_1, X_2) - SSR(X_2) \\
&= 1,814.4333 - 1,398.0345 \\
&= 416.3988
\end{aligned}
$$

คำนวณค่าสัมประสิทธิ์ตัวกำหนดบางส่วนได้ดังนี้

$$
\begin{aligned}
r^2_{Y1.2} &= \frac{SSR(X_1 \mid X_2)}{SSE(X_2)} \\
&= \frac{416.3988}{587.6988} \\
&= 0.7085 \\
r^2_{Y2.1} &= \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} \\
&= \frac{1,424.3769}{1,595.6769} \\
&= 0.8906
\end{aligned}
$$

นั่นคือ การเพิ่ม X_1 เข้าไปในสมการที่มี X_2 อยู่แล้ว ทำให้ค่า $SSE(X_2)$ ลดลง 70.85% และทำนองเดียวกัน การเพิ่ม X_2 เข้าไปในสมการถดถอยที่มี X_1 อยู่แล้ว ทำให้ค่า $SSE(X_1)$ ลดลง 89.06%

คำนวณค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนได้เป็น

$$r_{Y1.2} = \sqrt{r^2_{Y1.2}} = \sqrt{0.7085} = 0.8417$$
$$r_{Y2.1} = \sqrt{r^2_{Y2.1}} = \sqrt{0.8906} = 0.9437$$

3.10 โมเดลถดถอยเชิงเส้นตรงแบบพหุในรูปของคะแนนมาตรฐาน (Standardized Multiple Regression Model)

การเปรียบเทียบค่าสัมประสิทธิ์ถดถอยในทางปฏิบัติมักทำไม่ได้ เนื่องจากขนาดของค่าสัมประสิทธิ์ถดถอย b_j จะขึ้นอยู่กับหน่วยการวัดของตัวแปรอิสระ X_j พิจารณาจากสมการถดถอยต่อไปนี้

$$\hat{Y} = 150 + 3000X_1 + 3X_2$$

หาก Y มีหน่วยเป็นบาท X_1 มีหน่วยเป็นกิโลกรัม และ X_2 มีหน่วยเป็นกรัม ถึงแม้ว่า b_1 มีค่ามากกว่า b_2 มาก ๆ แต่ตัวแปรอิสระทั้งสองมีอิทธิพลต่อ Y เท่า ๆ กัน นั่นคือ ไม่ว่า X_1 หรือ X_2 เปลี่ยนแปลงไป 1 กิโลกรัม (เมื่อตัวแปรอิสระอื่นคงที่) ตัวแปรตาม Y จะเปลี่ยนแปลงไปในขนาดที่เท่ากัน ด้วยเหตุนี้การสร้างสมการถดถอยแสดงความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระที่ถูกสเกลค่าแล้ว จะทำให้ได้ค่าสัมประสิทธิ์ถดถอยที่อยู่ในหน่วยเดียวกัน และสามารถนำมาเปรียบเทียบกันได้ นอกจากนี้การสเกลค่าของตัวแปรอิสระยังเป็นการควบคุมความคลาดเคลื่อนที่เกิดจากการปัดเศษทศนิยม (Roundoff errors) ในขั้นตอนของการคำนวณ $(\mathbf{X}'\mathbf{X})^{-1}$ สำหรับสมการปกติ ซึ่งมักเกิดขึ้นในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง หรือตัวแปร X แต่ละตัวมีค่าแตกต่างกันมาก โดยวิธีที่นิยมใช้ในการสเกลมี 2 วิธีด้วยกัน ดังนี้

1. **วิธีสเกลปกติหนึ่งหน่วย (Unit normal scaling)** การสเกลค่าของตัวแปรอิสระและตัวแปรตามทำได้ดังนี้

$$Y_i^{(n)} = \frac{Y_i - \bar{Y}}{S_j}, \qquad i = 1, 2, \ldots, n \tag{3.45}$$
$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_y}, \qquad j = 1, 2, \ldots, k \tag{3.46}$$

เมื่อ

$S_j^2 = \dfrac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-1}$ แทน ความแปรปรวนของ X_j ในตัวอย่าง

$$S_y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$ แทน ความแปรปรวนของ Y ในตัวอย่าง

จะเห็นได้ว่าการสเกลค่าของตัวแปรอิสระและตัวแปรตามข้างต้นมีลักษณะคล้ายคลึงกับการแปลงตัวแปรสุ่มที่มีการแจกแจงแบบปกติให้อยู่ในรูปของคะแนนมาตรฐาน โดยตัวแปรที่ถูกสเกลค่าแล้วจะมีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1 นอกจากนี้การที่ตัวแปรตามและตัวแปรอิสระถูกหักออกด้วย $\bar{Y}$ และ $\bar{X}_j$ ตามลำดับ ส่งผลให้โมเดลถดถอยไม่มีระยะตัดแกน Y นั่นคือ $b'_0 = \bar{Y}^{(n)} = 0$ ดังนั้นโมเดลถดถอยสำหรับตัวแปรที่ถูกสเกลค่าแล้วคือ

$$Y_i^{(n)} = b'_1 Z_{i1} + b'_2 Z_{i2} + \cdots + b'_k Z_{ik} + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{3.47}$$

ซึ่งตัวประมาณกำลังสองน้อยที่สุดของพารามิเตอร์ β คือ

$$\mathrm{b}' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^{(n)} \tag{3.48}$$

2. **วิธีสเกลความยาวหนึ่งหน่วย (Unit length scaling)** การสเกลค่าของตัวแปรอิสระและตัวแปรตามทำได้ดังนี้

$$Y_i^{(0)} = \frac{Y_i - \bar{Y}}{\sqrt{S_{yy}}}, \qquad i = 1, 2, \ldots, n \tag{3.49}$$

$$W_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{S_{jj}}}, \qquad j = 1, 2, \ldots, k \tag{3.50}$$

เมื่อ

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \\ S_{jj} &= \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 = \sum_{i=1}^{n} X_{ij}^2 - n\bar{X}_j^2 \end{aligned}$$

การสเกลตัวแปรอิสระข้างต้น ทำให้ได้ว่าตัวแปรอิสระแต่ละตัว (W_j) มีค่าเฉลี่ยเป็น 0 และมีความยาวเป็น 1 นั่นคือ

$$\bar{W}_j = 0 \qquad \text{และ} \qquad \sqrt{\sum_{i=1}^{n}(W_{ij} - \bar{W}_j)^2} = 1$$

และได้โมเดลถดถอยดังนี้

$$Y_i^{(0)} = b_1' W_{i1} + b_2' W_{i2} + \cdots + b_k' W_{ik} + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{3.51}$$

โดยตัวประมาณกำลังสองน้อยที่สุดของพารามิเตอร์ β คือ

$$\mathbf{b}' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}^{(0)} \tag{3.52}$$

เมื่อ $\mathbf{W}'\mathbf{W}$ คือ เมตริกซ์ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation matrix) นั่นคือ

$$\mathbf{W}'\mathbf{W} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{bmatrix}$$

โดยที่ $r_{\ell j}$ แทน ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_ℓ และ X_j นั่นคือ

$$r_{\ell j} = \frac{S_{\ell j}}{\sqrt{S_{\ell\ell} \cdot S_{jj}}} = \frac{\sum_{i=1}^{n}(X_{i\ell} - \bar{X}_\ell)(X_{ij} - \bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_{i\ell} - \bar{X}_\ell)^2 \cdot \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}$$

และ $\mathbf{W}'\mathbf{Y}^{(0)} = \begin{bmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \\ \vdots \\ r_{yk} \end{bmatrix}$

โดยที่ r_{yj} แทน ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง Y และ X_j นั่นคือ

$$r_{yj} = \frac{S_{yj}}{\sqrt{S_{yy} \cdot S_{jj}}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{ij} - \bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \cdot \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}$$

เนื่องจาก $\mathbf{Z'Z}$ ที่ได้โดยวิธีสเกลปกติหนึ่งหน่วยมีความสัมพันธ์อย่างใกล้ชิดกับ $\mathbf{W'W}$ นั่นคือ

$$\mathbf{Z'Z} = (n-1)\mathbf{W'W}$$

ดังนั้นไม่ว่าจะใช้วิธีสเกลตัวแปรแบบใดก็ตาม ยังคงได้ค่าสัมประสิทธิ์ถดถอยค่าเดียวกัน นอกจากนี้ค่าสัมประสิทธิ์ถดถอยที่ได้จากวิธีปกติ $b_1, b_2, \ldots, b_k$ ยังมีความสัมพันธ์กับค่าสัมประสิทธิ์ถดถอยที่ได้จากสมการที่อยู่ในรูปของคะแนนมาตรฐาน $b'_1, b'_2, \ldots, b'_k$ ดังนี้

$$b_j = b'_j\sqrt{\frac{S_{yy}}{S_{jj}}}, \qquad j = 1, 2, \ldots, k \tag{3.53}$$

และ

$$b_0 = \bar{Y} - b_1\bar{X}_1 - \cdots - b_k\bar{X}_k \tag{3.54}$$

การแปลผลค่าสัมประสิทธิ์ถดถอยในสมการที่อยู่ในรูปคะแนนมาตรฐานยังสามารถทำได้ในลักษณะเช่นเดิม นั่นคือ b_j แสดงถึงอิทธิพลของ X_j เมื่อกำหนดให้ตัวแปรอิสระอื่นทั้งหมดอยู่ในสมการด้วย นอกจากนี้ขนาดของ b_j ยังขึ้นอยู่กับพิสัยของค่าของตัวแปรอิสระ ดังนั้นการใช้ขนาดของ b_j ในการเปรียบเทียบอิทธิพลของ X_j จึงควรทำด้วยความระมัดระวัง

ตัวอย่างที่ 3.10 จากข้อมูลในตัวอย่างที่ 3.1 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุที่อยู่ในรูปคะแนนมาตรฐานโดยใช้วิธีสเกลความยาวหนึ่งหน่วย

วิธีทำ จาก

$$\begin{array}{lll}
\sum_{i=1}^{n} X_{i1} = 192, & \sum_{i=1}^{n} X_{i2} = 76, & \sum_{i=1}^{n} Y_i = 598, \\
\sum_{i=1}^{n} X_{i1}^2 = 2,780, & \sum_{i=1}^{n} X_{i2}^2 = 524, & \sum_{i=1}^{n} X_{i1}X_{i2} = 953, \\
\sum_{i=1}^{n} X_{i1}Y_i = 8,009, & \sum_{i=1}^{n} X_{i2}Y_i = 3,451, & \sum_{i=1}^{n} Y_i^2 = 25,826,
\end{array}$$

$$n = 15, \qquad \bar{X}_1 = 12.8, \qquad \bar{X}_2 = 5.0667, \qquad \bar{Y} = 39.8667$$

คำนวณค่าต่าง ๆ ได้ดังนี้

$$
\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 25,826 - 15(39.8667)^2 = 1,985.7333 \\
S_{y1} &= \sum_{i=1}^{n} Y_i X_{i1} - n\bar{Y}\bar{X}_1 = 8,009 - 15(39.8667)(12.8) = 354.6 \\
S_{y2} &= \sum_{i=1}^{n} Y_i X_{i2} - n\bar{Y}\bar{X}_2 = 3,451 - 15(39.8667)(5.0667) = 421.1134 \\
S_{11} &= \sum_{i=1}^{n} X_{i1}^2 - n\bar{X}_1^2 = 2,780 - 15(12.8)^2 = 322.4 \\
S_{22} &= \sum_{i=1}^{n} X_{i2}^2 - n\bar{X}_2^2 = 524 - 15(5.0667)^2 = 138.9283 \\
S_{12} &= \sum_{i=1}^{n} X_{i1} X_{i2} - n\bar{X}_1\bar{X}_2 = 953 - 15(12.8)(5.0667) = -19.8064
\end{aligned}
$$

คำนวณค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทุกคู่ได้เป็น

$$
\begin{aligned}
r_{12} &= \frac{S_{12}}{\sqrt{S_{11} \cdot S_{22}}} = \frac{-19.8064}{\sqrt{(322.4)(138.9283)}} = -0.0936 \\
r_{y1} &= \frac{S_{y1}}{\sqrt{S_{yy} \cdot S_{11}}} = \frac{354.6}{\sqrt{(1985.7333)(322.4)}} = 0.4432 \\
r_{y2} &= \frac{S_{y2}}{\sqrt{S_{yy} \cdot S_{22}}} = \frac{421.1134}{\sqrt{(1985.7333)(138.9283)}} = 0.8018
\end{aligned}
$$

และได้ว่า $\mathbf{W'W} = \begin{bmatrix} 1 & -0.0936 \\ -0.0936 & 1 \end{bmatrix}, \qquad \mathbf{W'Y}^{(0)} = \begin{bmatrix} 0.4432 \\ 0.8018 \end{bmatrix}$

ดังนั้นค่าประมาณของสัมประสิทธิ์ถดถอยคือ

$$
\begin{aligned}
\mathbf{b}' &= (\mathbf{W'W})^{-1}\mathbf{W'Y}^{(0)} \\
&= \begin{bmatrix} 1 & -0.0936 \\ -0.0936 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.4432 \\ 0.8018 \end{bmatrix} \\
&= \begin{bmatrix} 1.00884 & 0.09443 \\ 0.09443 & 1.00884 \end{bmatrix} \begin{bmatrix} 0.4432 \\ 0.8018 \end{bmatrix} \\
&= \begin{bmatrix} 0.5228 \\ 0.8507 \end{bmatrix}
\end{aligned}
$$

และได้สมการถดถอยในรูปคะแนนมาตรฐานดังนี้

$$\hat{Y}^{(0)} = 0.5228W_1 + 0.8507W_2$$

นั่นคือ เมื่อให้ W_2 อยู่ในสมการ หากคะแนนมาตรฐาน W_1 เพิ่มขึ้น 1 หน่วย $\hat{Y}^{(0)}$ จะสูงขึ้น 0.5228 หน่วย และทำนองเดียวกัน เมื่อให้ W_1 อยู่ในสมการ หากคะแนนมาตรฐาน W_2 เพิ่มขึ้น 1 หน่วย $\hat{Y}^{(0)}$ จะสูงขึ้น 0.8507 หน่วย ดังนั้นจึงดูเหมือนว่าดัชนีการปรับตัวมีอิทธิพลต่อความพึงพอใจในการทำงานของพนักงานมากกว่าระดับสติปัญญา เนื่องจาก b_2' มีค่ามากกว่า b_1' แต่อย่างไรก็ตามการแปลผลค่าสัมประสิทธิ์ถดถอยที่อยู่ในรูปคะแนนมาตรฐานควรที่จะทำด้วยความระมัดระวังดังได้กล่าวไว้แล้ว เนื่องจากค่าดังกล่าวขึ้นอยู่กับลักษณะการกระจายของข้อมูล และไม่ควรที่จะใช้ขนาดของค่าสัมประสิทธิ์ถดถอยที่อยู่ในรูปคะแนนมาตรฐานเพียงอย่างเดียวในการเปรียบเทียบอิทธิพลของตัวแปรอิสระ หากต้องการแปลงให้อยู่ในรูปของตัวแปรเดิม จะได้ว่า

$$\begin{aligned}
b_1 &= b_1'\sqrt{\frac{S_{yy}}{S_{11}}} = 0.5228\sqrt{\frac{1,985.7333}{322.4}} = 1.2975 \\
b_2 &= b_2'\sqrt{\frac{S_{yy}}{S_{22}}} = 0.8507\sqrt{\frac{1,985.7333}{138.9283}} = 3.2162 \\
b_0 &= \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 39.8667 - (1.2975)(12.8) - (3.2162)(5.0667) = 6.9632
\end{aligned}$$

นั่นคือ

$$\hat{Y} = 6.9632 + 1.2975X_1 + 3.2162X_2$$

ซึ่งก็คือสมการถดถอยที่ได้ในตัวอย่างที่ 3.1 นั่นเอง

แบบฝึกหัดบทที่ 3

1. ในการหาศึกษาอิทธิพลของอุณหภูมิเฉลี่ยต่อชั่วโมง (X_1) และดัชนีความเย็น (X_2) ต่อปริมาณการใช้เชื้อเพลิง (Y) ได้ข้อมูลดังนี้

ค่าสังเกตที่	1	2	3	4	5	6	7	8
X_1	25.0	29.8	27.3	42.4	55.9	48.3	59.7	63.2
X_2	19	28	16	20	17	9	2	0
Y	12.5	13.1	12.0	10.7	8.9	9.3	8.8	8.6

1.1 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุ

1.2 ที่ระดับนัยสำคัญ 0.01 จงทดสอบนัยสำคัญของสมการถดถอยที่ได้ในข้อ 1.1

1.3 จงคำนวณค่าความคลาดเคลื่อนมาตรฐานของการประมาณ

1.4 จงคำนวณค่าสัมประสิทธิ์ของตัวกำหนดแบบพหุ รวมทั้งแปลผลค่าที่ได้

1.5 ที่ระดับนัยสำคัญ 0.01 จงทดสอบนัยสำคัญของ α พร้อมทั้งสรุปผล ที่ได้

1.6 ที่ระดับนัยสำคัญ 0.01 จงทดสอบอิทธิพลของอุณหภูมิและดัชนีความเย็น โดยใช้ t-test พร้อมทั้งสรุปผลที่ได้

1.7 จงหาช่วงความเชื่อมั่น 99% ของ β_0, β_1 และ β_2 พร้อมทั้งแปลผลค่าที่ได้

1.8 จงหาช่วงความเชื่อมั่น 99% ของการใช้เชื้อเพลิงเฉลี่ย สำหรับค่าสังเกตที่มีอุณหภูมิเฉลี่ยเป็น 40 และดัชนีความเย็นเป็น 10 พร้อมทั้งแปลผลค่าที่ได้

1.9 จงหาช่วงแห่งการพยากรณ์ 99% ของการใช้เชื้อเพลิง สำหรับค่าสังเกตที่มีอุณหภูมิเฉลี่ยเป็น 40 และดัชนีความเย็นเป็น 10 พร้อมทั้งแปลผลค่าที่ได้

1.10 ที่ระดับนัยสำคัญ 0.01 จงหาค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่างการใช้เชื้อเพลิงและดัชนีความเย็น เมื่อกำหนดให้อุณหภูมิมีค่าคงที่และอยู่ในโมเดลแล้ว

2. องค์การคุ้มครองสิ่งแวดล้อมต้องการสร้างสมการพยากรณ์การเผาผลาญเชื้อเพลิง (mpg) ของรถโดยสารของบริษัทแห่งหนึ่ง จึงเก็บรวบรวมข้อมูลจากความเร็วเฉลี่ยของรถ (mph) ความสูงของรถ (พันฟุต) และน้ำหนักของผู้โดยสาร (ร้อยปอนด์) ดังนี้

การเผาผลาญเชื้อเพลิง Y	ความเร็วเฉลี่ยของรถ X_1	ความสูงของรถ X_2	น้ำหนักของผู้โดยสาร X_3
34	46	4	3
32	51	2	4
34	40	3	3
24	42	7	5
37	53	2	3
31	49	4	4
30	50	6	4
27	56	6	3
29	55	3	6
32	48	3	5
31	40	4	2
33	46	3	4

2.1 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุ และทดสอบความมีนัยสำคัญของสมการดังกล่าว ที่ระดับนัยสำคัญ 0.05

2.2 จงคำนวณค่าสัมประสิทธิ์ของตัวกำหนดแบบพหุ พร้อมทั้งแปลผล

2.3 จงคำนวณค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่าง Y กับ X_1 เมื่อกำหนดให้ X_2 และ X_3 คงที่และอยู่ในสมการ

2.4 ที่ระดับนัยสำคัญ 0.05 จงทดสอบอิทธิพลของความเร็วเฉลี่ยต่อการเผาผลาญเชื้อเพลิง โดยใช้ t-test พร้อมทั้งสรุปผลที่ได้

2.5 สมมติว่าท่านสร้างสมการถดถอย โดยมีตัวแปรอิสระ X_2 และ X_3 อยู่ในสมการ ที่ระดับนัยสำคัญ 0.05 จงใช้ Partial-F test ทดสอบว่า ควรเพิ่มตัวแปร X_1 เข้าไปในสมการดังกล่าวหรือไม่ พร้อมทั้งระบุสมการถดถอยที่ท่านจะเลือกใช้ ในการพยากรณ์

2.6 จงใช้สมการที่กำหนดให้ในข้อ 2.5 เพื่อหาช่วงความเชื่อมั่น 95% ของการเผาผลาญเชื้อเพลิงเฉลี่ย และช่วงแห่งการพยากรณ์ 95% ของการเผาผลาญเชื้อเพลิง สำหรับค่าสังเกตที่มีความสูงของรถเป็น 3000 ฟุต และมีน้ำหนักผู้โดยสารเป็น 400 ปอนด์ พร้อมทั้งแปลผลค่าที่ได้

3. พนักงานตรวจสอบของโรงงานแห่งหนึ่งต้องการศึกษาอิทธิพลของอุณหภูมิ (เซลเซียส) และความดัน (ปอนด์ต่อตารางนิ้ว) ต่อความแข็งของภาชนะพลาสติก จึงสุ่มตัวอย่างภาชนะมา 10 รายการ ได้ข้อมูลดังนี้

รายการที่	1	2	3	4	5	6	7	8	9	10
ความแข็งของภาชนะ	30.0	28.0	29.2	15.4	20.1	33.1	26.6	27.0	32.8	18.6
อุณหภูมิ	285	277	234	208	256	271	211	217	248	206
ความดัน	11	13	12	22	18	13	13	11	15	23

3.1 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุ และทดสอบความมีนัยสำคัญของสมการดังกล่าว ที่ระดับนัยสำคัญ 0.05

3.2 จงคำนวณค่าสัมประสิทธิ์ของตัวกำหนดแบบพหุ พร้อมทั้งแปลผล

3.3 จงคำนวณค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างความแข็งของภาชนะกับอุณหภูมิ เมื่อกำหนดให้ความดันมีค่าคงที่และอยู่ในสมการ

3.4 ที่ระดับนัยสำคัญ 0.05 จงทดสอบอิทธิพลของอุณหภูมิและความดัน โดยใช้ t-test พร้อมทั้งสรุปผลที่ได้

3.5 ที่ระดับนัยสำคัญ 0.05 จงใช้ Partial-F test ทดสอบว่าควรเพิ่มความดันเข้าไปในสมการที่มีอุณหภูมิอยู่แล้วหรือไม่ เปรียบเทียบผลที่ได้กับข้อ 3.4

3.6 จงสร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายที่มีความดันเป็นตัวแปรอิสระเพียงตัวเดียว เพื่อหาช่วงความเชื่อมั่น 95% ของความแข็งเฉลี่ย และช่วงแห่งการพยากรณ์ 95% ของความแข็งของพลาสติกสำหรับค่าสังเกตที่มีความดันเป็น 18 ปอนด์ต่อตารางนิ้ว พร้อมทั้งแปลผล

4. ผู้จัดการโรงพยาบาลแห่งหนึ่งต้องการศึกษาความสัมพันธ์ระหว่างความพอใจของคนไข้ (Y) กับอายุของคนไข้ (X_1) หน่วยเป็นปี ความรุนแรงของโรค (X_2) หน่วยเป็นเลขดัชนี และระดับความกังวล (X_3) หน่วยเป็นเลขดัชนี จึงได้เก็บรวบรวมข้อมูลจากคนไข้ทั้งหมด 15 คน สร้างตาราง ANOVA สำหรับ Full model และ Reduced model ดังนี้

a. สมการ $\hat{Y} = 6.564 + 2.223X_2$

Source of variation	df	SS	MS	F	P − value
Regression	1	8.196	8.196	4.623	0.051
Error	13	23.049	1.773		
Total	14	31.245			

b. สมการ $\hat{Y} = 1.621 + 2.494X_2 + 0.269X_3$

Source of variation	df	SS	MS	F	P − value
Regression	2	16.607	8.304	6.807	0.011
Error	12	14.638	1.220		
Total	14	31.245			

c. สมการ $\hat{Y} = 0.258 + 0.460X_1 + 1.646X_2 + 0.066X_3$

Source of variation	df	SS	MS	F	P − value
Regression	3	26.627	8.876	21.140	0.000
Error	11	4.618	0.420		
Total	14	31.245			

จากข้อมูลข้างต้น ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่า

4.1 ตัวแปรอิสระทั้ง 3 ตัวรวมกัน มีอิทธิพลต่อตัวแปรตาม Y หรือไม่ ที่ระดับนัยสำคัญ 0.05

4.2 เมื่อกำหนดให้ X_2 อยู่ในสมการแล้ว ตัวแปร X_1 และ X_3 สมควรอยู่ในสมการหรือไม่

4.3 เมื่อกำหนดให้ X_2 และ X_3 อยู่ในสมการแล้ว ตัวแปร X_1 สมควรอยู่ในสมการหรือไม่

4.4 จงคำนวณค่าสัมประสิทธิ์ตัวกำหนดบางส่วนระหว่าง Y กับ X_1 เมื่อกำหนดให้ X_2 และ X_3 มีค่าคงที่

4.5 จงเขียน Full model และ Reduced model ของสมการในข้อ 4.3