บทที่ ๒

ทฤษฎีที่ใช้ในการวิเคราะห์ปัญหา

ในการหาความสัมพันธ์ของข้อมูล ๒ พวก โดยที่พวกหนึ่งเป็นตัวแปรอิสระ (independent-variable) คือตัวแปรที่มีการเปลี่ยนแปลงค่าไม่ขึ้นกับตัวแปรอื่น กับอีกพวกหนึ่งเป็นตัวแปรตาม (dependent variable) ซึ่งหมายถึงตัวแปรที่มีการเปลี่ยนแปลงค่าไปตามตัวแปรอิสระ และอาจเป็นการเปลี่ยนแปลงไปในทางเดียวกันหรือในทางกลับกันก็ได้นั้น เราใช้ทฤษฎีความถดถอย (Regression Theory) เป็นเครื่องมือในการพิจารณา ซึ่งนอกจากจะได้ทราบถึงความสัมพันธ์ระหว่างข้อมูลสองพวกนั้นแล้ว ยังสามารถใช้สมการถดถอยทำนายค่าของตัวแปรตาม โดยอาศัยค่าของตัวแปรอิสระเป็นตัวทำนายได้อีกด้วย สำหรับข้อมูลในปัญหาที่จะทำการวิจัยนี้ เมื่อได้ลองนำไปเขียน แผนภาพการกระจาย (scatter diagram) ของตัวแปรอิสระกับตัวแปรตามแล้วพบว่า ความสัมพันธ์ของตัวแปรทั้งสองพวกจัดเป็นความสัมพันธ์ในเชิงเส้นตรง ดังนั้นทฤษฎีที่จะใช้ในการวิเคราะห์ปัญหานี้ จึงใช้ทฤษฎีความถดถอยในรูปของสมการเส้นตรง (Linear Regression Theory)(๒) ซึ่งมีแบบจำลองทางคณิตศาสตร์ (Mathematical Model)ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{-----} + \beta_p X_p + e \qquad (1)$$

เมื่อ Y เป็นตัวแปรตาม

X_i เป็นตัวแปรอิสระตัวที่ i , ($i = 1,2,3,\ldots,p$)

β_0 เป็นจุดตัดบนแกน Y ของเส้นถดถอย (regression line) หรือเป็นค่าเฉลี่ยของ Y ซึ่งแยกเอาความเปลี่ยนแปลงของค่า Y เนื่องจากค่า X_i ต่าง ๆ ออกไปแล้ว

β_i เป็นค่าสัมประสิทธิ์ของความถดถอย (regression coefficient) ($i = 1,2,3,\ldots,p$)ซึ่งแสดงถึง อัตราการเปลี่ยนแปลงของ Y เมื่อเทียบกับตัวแปรอิสระตัวที่ i ในขณะที่กำหนดให้ตัวแปรอิสระตัวอื่น ๆ คงที่

---

(๒) Draper & Smith , Applied Regression Analysis ,John Wiley & Sons Inc.

e เป็นความคลาดเคลื่อนซึ่งเนื่องจากการที่ ค่าประมาณของตัวแปรตาม Y สำหรับตัวแปรอิสระแต่ละชุด ที่ได้จากสมการถดถอย แตกต่างไปจากค่าจริง มีลักษณะเป็นตัวแปรสุ่ม (random variable) อาจเขียนได้ว่า $e = Y - \hat{Y}$

เมื่อ $\hat{Y}$ เป็นค่าประมาณของตัวแปรตาม Y ที่ได้จากสมการถดถอย โดยมีข้อสมมติว่า $E(e) = 0$; $V(e) = V(Y) = \sigma^2$; $COV(e_i , e_k) = 0$, $i \neq k$;

$$e \sim N (0 , \sigma^2)$$

จาก (1) ถ้าเราจะใช้ค่าตัวแปรอิสระต่าง ๆ ทำนายค่าตัวแปรตาม จะได้สมการเป็น

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \text{---------} + b_pX_p \qquad (2)$$

เมื่อ $\hat{Y}$ เป็นค่าประมาณของ Y สำหรับ X_i แต่ละชุด

b_0 เป็นค่าประมาณของ β_0

b_i เป็นค่าประมาณของ β_i , ($i = 1,2,3,\ldots\ldots, p$)

การคำนวณค่าจุดตัดบนแกน Y (b_0) และค่าสัมประสิทธิ์ของความถดถอย (b_i)

ในการคำนวณค่า b_0 และค่าสัมประสิทธิ์ของความถดถอย $b_i (i = 1,2,\ldots.p$) เพื่อให้ค่าที่ได้ออกมา เมื่อนำไปใช้ในสมการถดถอยแล้วทำให้ค่าประมาณของ Y ที่ได้ มีความคลาดเคลื่อนน้อยที่สุด วิธีคำนวณจึงใช้วิธีกำลังสองน้อยที่สุด (Method of Least Squaros) โดยให้ผลบวกของกำลังสองของความคลาดเคลื่อน มีค่าน้อยที่สุด

จาก (1) เมื่อมีข้อมูลจำนวน n ชุด ($j = 1,2,\ldots.,n$) จะได้สมการเป็น

$$Y_j = \beta_0 + \beta_1X_{1j} + \beta_2X_{2j} + \text{---------} + \beta_pX_{pj} + e_j \qquad (3)$$

$$\therefore e_j = Y_j - \beta_0 - \beta_1X_{1j} - \beta_2X_{2j} - \text{---------} - \beta_pX_{pj}$$

เมื่อรวมกำลังสองของความคลาดเคลื่อนทั้งชุดมูล จะได้เป็น

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1X_{1j} - \beta_2X_{2j} - \text{---------} - \beta_pX_{pj})^2 \qquad (4)$$

จาก (4) ทำการแก้สมการโดยอาศัยวิธี พาเชียลดีริเวทีฟ (partial derivative) เทียบกับ β_i (i = 1,2,....,p) และ β_0 แล้วเทียบให้เท่ากับ o ซึ่งจะได้สมการทั้งหมด (p+1) สมการ จากสมการเหล่านี้ เมื่อแก้สมการแล้วก็จะได้ค่าประมาณของ β_0 และ β_i ออกมาคือ b_0 , b_i (i = 1,2,3,.....,p)

จากสมการซึ่งทำ พาเชียลดีริเวทีฟ เทียบกับ β_0 และให้เท่ากับ o จะทำให้ได้ค่า b_0 ดังต่อไปนี้

$$\frac{\partial(\sum_{j=1}^{n} e_j^2)}{\partial b_0} = -2\sum_{j=1}^{n}(Y_j - b_0 - b_1X_{1j} - b_2X_{2j} - \text{---} - b_pX_{pj})^2 = 0$$

$$\therefore \sum_{j=1}^{n} b_0 = \sum_{j=1}^{n}(Y_j - b_1X_{1j} - b_2X_{2j} - \text{---} - b_pX_{pj})$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \text{---} - b_p\bar{X}_p \qquad (5)$$

จะเห็นได้ว่าค่า b_0 เป็นฟังก์ชั่นของ b_i, (i = 1,2,..p) ด้วยเหตุนี้ถ้าเราแทนค่า b_0 ใน (4) เสียก่อนที่จะคำนวณ พาเชียลดีริเวทีฟ โดยเทียบกับ b_i, (i = 1,2,..p) แต่ละตัวต่อไป ก็จะทำให้เหลือค่าที่จะต้องคำนวณเพียง P ค่า แทนที่จะเป็น P+1 ค่า ดังนี้

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n}\left[Y_j - \bar{Y} - b_1(X_{1j}-\bar{X}_1) - b_2(X_{2j}-\bar{X}_2) - \text{---} - b_p(X_{pj}-\bar{X}_p)\right]^2$$

เมื่อให้ $y = (Y_j - \bar{Y})$, $x_1 = (X_{1j} - \bar{X}_1)$, $x_2 = (X_{2j} - \bar{X}_2)$,.....

....., $x_p = (X_{pj} - \bar{X}_p)$

ก็จะสามารถเขียนสมการใหม่ได้ดังนี้

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y - b_1x_1 - b_2x_2 - \text{----} - b_px_p)^2 \qquad (6)$$

จากนั้นก็ทำ พาเชียลดีริเวทีฟ ของสมการ (6) เทียบต่อ b_i , (i = 1,2,..,p) ทีละตัว จะได้ P สมการ ซึ่งเมื่อแก้สมการเหล่านี้ก็จะได้ค่า b_1, b_2,..b_p ออกมา และนำไปหาค่า b_0 ได้

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = 0 \;,\; b_1 \Sigma(x_1^2) + b_2 \Sigma(x_1 x_2) + \cdots + b_p \Sigma(x_1 x_p) = \Sigma(x_1 y)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = 0 \;,\; b_1 \Sigma(x_2 x_1) + b_2 \Sigma(x_2^2) + \cdots + b_p \Sigma(x_2 x_p) = \Sigma(x_2 y)$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_p} = 0 \;,\; b_1 \Sigma(x_p x_1) + b_2 \Sigma(x_p x_2) + \cdots + b_p \Sigma(x_p^2) = \Sigma(x_p y) \qquad (7)$$

สมการ (7) เรียกว่า Normal Equations สำหรับตัวแปรอิสระ P ตัว ค่า b_o และ b_i ที่ได้จากการคำนวณโดยวิธีนี้ เป็นค่าประมาณของ β_o และ β_i ที่ไม่มีความลำเอียง (unbiased) และนอกจากนี้ความคลาดเคลื่อนมาตรฐาน (standard - error) ก็จะน้อยที่สุด เมื่อเปรียบเทียบกับค่าประมาณที่ไม่มีความลำเอียงอื่น ๆ ซึ่งอาจหาได้จากข้อมูลเดียวกัน

จาก (7) จะเห็นได้ว่ายิ่งมีตัวแปรอิสระมากขึ้นเท่าใด จำนวนสมการใน normal equation ก็จะยิ่งมากขึ้นเท่านั้น ซึ่งทำให้เกิดความซับซ้อนยุ่งยากในการแก้สมการเพื่อหาค่า b_i ออกมาทุก ๆ ตัว ดังนั้นเพื่อให้การแก้สมการหลาย ๆ สมการดังกล่าวแล้ว ทำได้ง่ายขึ้น เราอาจใช้วิธีการทาง แมทริกซ์ (Matrix) เข้าช่วยดังนี้

ความถดถอยเชิงซ้อนซึ่งเป็นเส้นตรง (Multiple Linear Regression) ในรูปแมทริกซ์

จากข้อมูล n ชุด เราสามารถจัดรูปของข้อมูลเป็นลักษณะของแมทริกซ์ได้ดังนี้

$$\begin{aligned}
Y_1 &= \beta_o + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_p X_{p1} + e_1 \\
Y_2 &= \beta_o + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_p X_{p2} + e_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
Y_n &= \beta_o + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_p X_{pn} + e_n
\end{aligned} \qquad (8)$$

เมื่อจัดเป็นรูปแมทริกซ์จะได้สมการเป็น

$$Y = X\beta + e \qquad (9)$$

เมื่อ

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n\times 1)} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix}_{(n\times(p+1))}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{(n\times 1)}$$

โดยมีข้อสมมติเบื้องต้นดังนี้

$E(e) = (0)$, $V(e) = V(Y) = I\sigma^2$, $e \sim N(\ (0)\ ,\ I\sigma^2\ )$

และสำหรับค่าประมาณของ Y คือ $\hat{Y}$ ก็จะมีสมการเป็น

$$\hat{Y} = Xb \qquad (10)$$

เมื่อ

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}_{(n\times 1)} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}_{[(p+1)\times 1]}$$

เมื่อพิจารณาจาก (10) จะได้ normal equations เพื่อนำไปหาค่า b ดังนี้

$$X'Xb = X'Y$$

$$[X'X]^{-1}X'Xb = [X'X]^{-1}X'Y$$

$$b_{[(p+1)\times 1]} = [X'X]^{-1}_{(p+1)(p+1)}\ X'Y_{[(p+1)\times 1]} \qquad (11)$$

จากสมการ (11) เราจะได้ค่า $b_0, b_1, \ldots, b_p$ ออกมา จะเห็นได้ว่าจาก normal equations ในสมการ (7) เราได้ค่า $b_1, b_2, \ldots, b_p$ เท่านั้น ค่า b_0 ต้องคำนวณใหม่

อีกครั้ง ดังนั้นการหาค่าในรูปแมทริกซ์ จะให้ผลดีกว่าการแก้สมการ normal equations โดยตรง เพราะสามารถหาค่า b_o ออกมาได้พร้อมกับค่า b_i แต่วิธีแมทริกซ์นี้จะใช้ได้เมื่อ iff $(X'X)$ เป็น nonsingular matrix

$$\left[X'X\right]_{(P+1)(P+1)} = \begin{bmatrix} n & \Sigma X_{1j} & \Sigma X_{2j} & \cdots & \Sigma X_{pj} \\ \Sigma X_{1j} & \Sigma X_{1j}^2 & \Sigma(X_{1j}X_{2j}) & \cdots & \Sigma(X_{1j}X_{pj}) \\ \Sigma X_{2j} & \Sigma(X_{2j}X_{1j}) & \Sigma X_{2j}^2 & \cdots & \Sigma(X_{2j}X_{pj}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma X_{pj} & \Sigma(X_{pj}X_{1j}) & \Sigma(X_{pj}X_{2j}) & & \Sigma X_{pj}^2 \end{bmatrix}_{(P+1)(P+1)}$$

$$\left[X'Y\right]_{[(P+1)\times 1]} = \begin{bmatrix} \Sigma Y_j \\ \Sigma(X_{1j}Y_j) \\ \Sigma(X_{2j}Y_j) \\ \cdots \\ \Sigma(X_{pj}Y_j) \end{bmatrix}_{[(p+1)\times 1]}$$

การหา Inverse Matrix $\left[X'X\right]^{-1}$ ใช้วิธี Abbreviated Doolittle (ดูจากภาคผนวก)

$\left[X'X\right]^{-1}_{(P+1)(P+1)}$ จะเป็น symmetric matrix ด้วย

$$\left[X'X\right]^{-1} = C = \begin{bmatrix} c_{oo} & c_{o1} & c_{o2} & \cdots & c_{op} \\ c_{1o} & c_{11} & c_{12} & \cdots & c_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{po} & c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}$$

ดังนั้นจะได้

$$b_o = c_{oo}\Sigma Y_j + c_{o1}\Sigma(X_{1j}Y_j) + c_{o2}\Sigma(X_{2j}Y_j) + \cdots + c_{op}\Sigma(X_{pj}Y_j)$$

$$b_1 = c_{1o}\Sigma Y_j + c_{11}\Sigma(X_{1j}Y_j) + c_{12}\Sigma(X_{2j}Y_j) + \cdots + c_{1p}\Sigma(X_{pj}Y_j)$$

$$b_2 = c_{2o}\Sigma Y_j + c_{21}\Sigma(X_{1j}Y_j) + c_{22}\Sigma(X_{2j}Y_j) + \cdots + c_{2p}\Sigma(X_{pj}Y_j)$$

$$\cdots\cdots\cdots\cdots$$

$$b_p = c_{po}\Sigma Y_j + c_{p1}\Sigma(X_{1j}Y_j) + c_{p2}\Sigma(X_{2j}Y_j) + \cdots + c_{pp}\Sigma(X_{pj}Y_j)$$

และ $V(b) = [X'X]^{-1} \sigma^2$ เมื่อ $V(b)$ เป็น Variance Covariance Matrix ของ b
นั่นคือ $s^2_{b_i} = c_{ii} s^2$ เมื่อ s^2 คือค่าประมาณของ σ^2 , $V(Y) = I\sigma^2$

การวิเคราะห์ข้อมูลโดยใช้ทฤษฎีความถดถอยเชิงซ้อนซึ่งเป็นเส้นตรง

ในการทดสอบสมมติฐานเพื่อหาความสัมพันธ์ของข้อมูล ๒ พวก คือระหว่างตัวแปรตามกับตัวแปรอิสระ โดยทฤษฎีความถดถอยเชิงซ้อนซึ่งเป็นเส้นตรงนี้ วิธีหนึ่งที่นิยมใช้ก็คือ การวิเคราะห์ความแปรปรวน (Analysis of Variance) ซึ่งใช้ค่า F เป็นเครื่องพิจารณา

ตามทฤษฎีนี้ เรากำหนดข้อสมมติว่า ค่าความคลาดเคลื่อน e มีการกระจายแบบโค้งปกติ (normal curve) มีค่าเฉลี่ย (mean) = 0 และความแปรปรวน (variance) = σ^2 นั่นคือ $e \sim N(0, \sigma^2)$
ค่าประมาณที่ไม่มีความลำเอียงของ σ^2 คือ $s^2 = \dfrac{\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2}{[n - (p+1)]}$

เมื่อ n เป็นจำนวนข้อมูล

p + 1 เป็นจำนวน พารามิเตอร์ (parameters) ที่ใช้ในแบบจำลอง

เนื่องจากวิธีที่ใช้ในการหาค่า $b_0, b_1, b_2, \ldots\ldots, b_p$ ของสมการถดถอยเราใช้วิธีผลบวกของกำลังสองของความคลาดเคลื่อนน้อยที่สุด ดังนั้นในการพิจารณาวิเคราะห์เพื่อทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เราจึงต้องพิจารณาจากค่า ผลบวกของกำลังสองของความคลาดเคลื่อนดังนี้

จาก $e_j = (Y_j - \hat{Y}_j) = (Y_j - \bar{Y}) - (\hat{Y}_j - \bar{Y})$

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} (Y_j - \bar{Y})^2 + \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 - 2\sum_{j=1}^{n} (Y_j - \bar{Y})(\hat{Y}_j - \bar{Y}) \quad (12)$$

เพื่อให้เห็นได้ชัดจะพิจารณากรณีที่มีตัวแปรอิสระ X_1 เพียงตัวเดียวในแบบจำลอง

จาก (12) จะเห็นได้ว่าค่า

$$\begin{aligned}
\sum_{j=1}^{n} (Y_j - \bar{Y})(\hat{Y}_j - \bar{Y}) &= \sum_{j=1}^{n} (Y_j - \bar{Y}) \left[\bar{Y} + b_1 (X_{1j} - \bar{X}_1) - \bar{Y} \right] \\
&= \sum_{j=1}^{n} (Y_j - \bar{Y}) \left[b_1 (X_{1j} - \bar{X}_1) \right] \\
&= b_1 \sum_{j=1}^{n} (Y_j - \bar{Y})(X_j - \bar{X}) \\
&= b_1 \left(b_1 \sum_{j=1}^{n} (X_j - \bar{X})^2 \right) \qquad b_1 = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}
\end{aligned}$$

$$= b_1^2 \sum_{j=1}^{n} (X_j - \bar{X})^2 = \sum_{j=1}^{n} \left[b_1 (X_j - \bar{X}) \right]^2$$

$$= \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y}) \quad (\text{เพราะ } b_1 (X_j - \bar{X}) = \hat{Y}_j - \bar{Y})$$

เมื่อนำไปแทนค่าใน (12)จะได้ดังนี้

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} (Y_j - \bar{Y})^2 + \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 - 2\sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2$$

$$= \sum_{j=1}^{n} (Y_j - \bar{Y})^2 - \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2$$

$$\therefore \sum_{j=1}^{n} (Y_j - \bar{Y})^2 = \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 + \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$$

นั่นคือ ผลบวกกำลังสองของความคลาดเคลื่อนเนื่องจากค่าจริงของตัวแปรตาม แตกต่างไปจากค่าเฉลี่ย (Total sum of Squares $= \sum_{j=1}^{n} (Y_j - \bar{Y})^2$) เท่ากับ ผลบวกของกำลังสองของความคลาดเคลื่อนเนื่องจากค่าประมาณของตัวแปรตามแตกต่างไปจากค่าเฉลี่ย (คือความเบี่ยงเบนคิดจากเส้นถดถอยกับค่าเฉลี่ย $= \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2$) รวมกับ ผลบวกของกำลังสองของความคลาดเคลื่อนเนื่องจากค่าจริงของตัวแปรตามแตกต่างไปจากค่าประมาณของตัวเอง (sum of squares - about regression หรือ Residual sum of Squares $= \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$)

ค่าต่าง ๆ ดังกล่าวนี้ถ้าเราพิจารณาข้อมูลในรูปแมทริกซ์ จะได้เป็นรูปดังนี้

$$\sum_{j=1}^{n} (Y_j - \bar{Y})^2 = (Y'Y)_{1\times 1}$$

$$\sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 = (b'X'Y)_{1\times 1}$$

$$\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = Y'Y - b'X'Y$$

จากค่าผลบวกกำลังสองทั้งสามค่านี้ เรานำไปวิเคราะห์ความแปรปรวนเพื่อทดสอบสมมติฐานของความสัมพันธ์ต่อไป

เนื่องจากค่าเฉลี่ยที่เรานำมาใช้มิใช่ค่าเฉลี่ยที่แท้จริงของข้อมูลในประชากร แต่เป็นค่าที่เราประมาณออกมาจากข้อมูลที่เป็นตัวอย่างในการพิจารณา เพื่อมิให้เกิดความลำเอียงในการคำนวณค่าต่าง ๆ จึงต้องมีการแก้ไข ดังแสดงไว้ในตารางวิเคราะห์ความแปรปรวนต่อไปนี้

ตารางวิเคราะห์ความแปรปรวน (ANOVA)

แหล่งความแปรปรวน S.V.	D.F.	SUM SQUARES	MEAN SQUARES	F	R^2 %
Due to Regression	p	$\hat{b}'X'Y - n\bar{Y}^2 = SSR$	$\frac{SSR}{p} = A$	A/B	$\frac{SSR}{SST} \times 100$
Residual(error)	n-p-1	$Y'Y - \hat{b}'X'Y = SSE$	$\frac{SSE}{n-p-1} = B$		
Total(corrected)	n-1	$Y'Y - n\bar{Y}^2$			

จะเห็นได้ว่าในการวิเคราะห์ความแปรปรวน เราอาจพิจารณาจากค่า F ประกอบกับค่า **สัมประสิทธิ์แห่งการตกลงใจ(coefficient of determination)** สำหรับค่า F เป็นค่าที่ใช้ในการทดสอบสมมติฐาน (null hypothesis) $H_0 : \beta_i = 0$ หรือ $H_0 : \beta_0 = \beta_1 = \beta_i = 0$ การหาค่า F พิจารณาจากอัตราส่วนของ ผลบวกกำลังสองของความแตกต่างระหว่างค่าจริงของตัวแปรตามที่แตกต่างไปจากค่าเฉลี่ยของมัน กับ ผลบวกกำลังสองของความแตกต่างระหว่างค่าจริงกับค่าประมาณของตัวแปรตาม โดยที่ได้นำค่า degrees of freedom ของค่านั้น ๆ ไปหารแล้ว สำหรับค่า degrees of freedom ของค่า F กำหนดด้วยค่า degrees of freedom ของค่าผลบวกกำลังสองของความแตกต่างทั้งสองค่าที่นำมาใช้คำนวณหาค่า F นั้น

$$F = \frac{\text{Mean Squares due to Regression}\ (D.F. = p)}{\text{Residual Mean Squares}\ (D.F. = n - p - 1)}$$

ค่า degrees of freedom ของ F คือ (p , n - p - 1)

จากค่า F ที่คำนวณได้ เรานำไปเปรียบเทียบกับค่า F จากตาราง ซึ่งมี degrees of freedom เดียวกัน ณ ระดับนัยสำคัญทางสถิติ ที่ต้องการพิจารณา เช่น ๐.๐๑ หรือ ๐.๐๕ ซึ่งใช้สัญญลักษณ์ α สำหรับในการวิเคราะห์ของปัญหานี้ จะใช้ระดับนัยสำคัญ $\alpha =$ ๐.๐๑

การเปรียบเทียบ ถ้าค่า $F_{(p,n-p-1)}$ ที่คำนวณได้ มากกว่า ค่า $F_{(p,n-p-1)}$ จากตาราง เราก็ต้อง ปฏิเสธสมมติฐาน null Hypothesis

ถ้าค่า $F_{(p,n-p-1)}$ ที่คำนวณได้ น้อยกว่า ค่า $F_{(p,n-p-1)}$ จากตาราง เราจะไม่ปฏิเสธสมมติฐาน null Hypothesis

ค่า Residual Mean Square ในตารางวิเคราะห์ความแปรปรวน (ANOVA) นั้น ก็คือค่า กำลังสองของค่าประมาณของความคลาดเคลื่อนมาตรฐาน (standard error) ที่เกิดขึ้นในการประมาณค่า ตัวแปรตาม Y ด้วยค่าตัวแปรอิสระ X_i ต่าง ๆ ในแบบจำลองนั้น

จาก $$s_{Y.1\ldots p} = \sqrt{\frac{\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{n-p-1}}$$

เมื่อ $s_{Y.1\ldots p}$ คือค่าประมาณของความคลาดเคลื่อนมาตรฐานในการประมาณค่าตัวแปรตาม Y ดังนั้นถ้าพิจารณาจากตารางวิเคราะห์ความแปรปรวนจะได้

$$s_{Y.1\ldots p} = \sqrt{\text{residual mean square}}$$

สำหรับค่าประมาณของความคลาดเคลื่อนมาตรฐานในการประมาณค่า $b_1, b_2, \ldots, b_p$ พิจารณาจาก Variance Covariance Matrix ของ b, $V(b) = V = [X'X]^{-1}\sigma^2 = C\sigma^2$

$$C = \begin{bmatrix} c_{oo} & c_{o1} & c_{o2} & \cdots & c_{op} \\ c_{1o} & c_{11} & c_{12} & \cdots & c_{1p} \\ c_{2o} & c_{21} & c_{22} & \cdots & c_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{po} & c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}_{(P+1)(P+1)}$$

$$s_{bi} = \sqrt{c_{ii}\sigma^2} = s_{Y.1\ldots p}\sqrt{c_{ii}}$$

เมื่อ s_{bi} เป็นค่าประมาณของความคลาดเคลื่อนมาตรฐานในการประมาณค่า b_i

c_{ii} เป็นค่าบนแนวทแยงมุม (diagonal) ของ แมทริกซ์ C ตัวที่ ii

($i = 0,1,2,\ldots,p$)

σ^2 เป็นความแปรปรวนของข้อมูลที่นำมาใช้เป็นตัวแปรตาม

นอกจากนี้ยังอาจหาค่าความแปรปรวนร่วม (covariance) ระหว่าง b_i กับ b_k ได้ดังนี้

$$COV(b_i, b_k) = \frac{c_{ik}\sigma^2}{\sqrt{c_{ii}\sigma^2 c_{kk}\sigma^2}}$$

c_{ik} เป็นค่าในแมทริกซ์ C ตัวที่ ik

($k = 0,1,2,\ldots,p$)

ค่าสหสัมพันธ์เชิงเดี่ยว (Simple Correlation)

ในกรณีที่เราต้องการทราบความสัมพันธ์ระหว่างข้อมูล ๒ พวก ว่ามีความสัมพันธ์ต่อกันมากน้อย เป็นไปในทางเดียวกันหรือกลับกันอย่างไร เราสามารถพิจารณาได้จากค่า สหสัมพันธ์ ระหว่างข้อมูลทั้ง ๒ ค่าสหสัมพันธ์ระหว่างข้อมูลที่มีตัวแปรตาม และตัวแปรอิสระ เพียงตัวเดียว เราเรียกว่าค่า สหสัมพันธ์เชิงเดี่ยว หรือ สหสัมพันธ์อย่างง่าย ซึ่งมีสัญญลักษณ์ r ถ้าต้องการหาค่าสหสัมพันธ์เชิงเดี่ยวระหว่างข้อมูล X และ Y เราจะใช้สัญญลักษณ์ r_{XY} ซึ่งหาได้ดังนี้

$$r = \frac{COV(X, Y)}{\sqrt{V(X)\,V(Y)}}$$

$$= \frac{\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2 \sum_{j=1}^{n}(Y_j - \bar{Y})^2}}$$

ค่า สหสัมพันธ์ จะมีค่าอยู่ระหว่าง -1 ถึง $+1$

ถ้าค่าสหสัมพันธ์เข้าใกล้ -1 แสดงว่า Y มีความสัมพันธ์กับ X มาก แต่เป็นไปในทางกลับกัน คือถ้า X มีค่าเพิ่มขึ้น Y จะมีค่าลดลง

ถ้าค่าสหสัมพันธ์ใกล้ $+1$ แสดงว่า Y มีความสัมพันธ์กับ X มาก และเป็นไปในทางเดียวกัน คือถ้า X มีค่าเพิ่มขึ้น Y จะมีค่าเพิ่มขึ้นด้วย

กล่าวโดยสรุปได้ว่า ถ้าค่าสัมบูรณ์ของ r, $|r|$ มีค่าใกล้ 1 แสดงว่า Y กับ X มีความสัมพันธ์ต่อกันสูง

ถ้าค่าสัมบูรณ์ของ r, $|r|$ มีค่าใกล้ ๐ แสดงว่า Y กับ X มีความสัมพันธ์กันน้อยมาก

ค่าสหสัมพันธ์เชิงซ้อน (Multiple Correlation)

เป็นค่าสหสัมพันธ์ระหว่างตัวแปรตาม Y กับตัวแปรอิสระ $X_1, X_2, \ldots, X_p$ ทั้งหมดที่นำมาใช้ในแบบจำลองความถดถอย ใช้สัญญลักษณ์ R เป็นค่าที่บอกให้ทราบว่า ตัวแปรตาม Y มีความสัมพันธ์ กับ ตัวแปรอิสระทุกตัวที่นำมาใช้ในแบบจำลองความถดถอย มากหรือน้อยเพียงใด (ไม่ใช่ค่าสหสัมพันธ์ เฉพาะกับตัวแปรอิสระตัวใดตัวหนึ่ง แต่เป็นค่าสหสัมพันธ์กับตัวแปรอิสระทั้งหมดทุกตัวในแบบจำลอง) หรือบอกให้ทราบถึงความสัมพันธ์ ระหว่าง ค่าตัวแปรตามจริง กับค่าประมาณของตัวแปรตาม ซึ่งคำนวณได้จากแบบจำลองความถดถอยนั้น จึงอาจเขียนได้ดังนี้

$$R = r_{Y\hat{Y}}$$

$$= \frac{COV(Y, \hat{Y})}{\sqrt{V(Y)\, V(\hat{Y})}}$$

ค่า R มีค่าอยู่ระหว่าง -1 ถึง $+1$ ความหมายของค่า R ที่แสดงถึงความสัมพันธ์ของตัวแปรตามกับตัวแปรอิสระ พิจารณาได้เช่นเดียวกับค่าสหสัมพันธ์เชิงเดี่ยว

ค่า R^2 เราเรียกว่า สัมประสิทธิ์แห่งการตกลงใจ (coefficient of determination) ซึ่งจะบอกให้ทราบว่า สัดส่วนของความแปรปรวนของตัวแปรตาม Y ที่ขึ้นกับความแปรปรวนของค่าตัวแปรอิสระ X_i ต่าง ๆ ในแบบจำลอง มีมากน้อยอย่างไร เมื่อเทียบกับค่าความแปรปรวนทั้งหมดของตัวแปรตาม Y นั้น

พิจารณาจากตารางวิเคราะห์ความแปรปรวน

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

$$= \frac{\sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}$$

หรือ $$= \frac{b'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$ (ในรูปแมทริกซ์)

เพราะว่า $\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} y_j^2 - R^2 \sum_{j=1}^{n} y_j^2$

$$\therefore \sum_{j=1}^{n} y_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 + R^2 \sum_{j=1}^{n} y_j^2$$

$$\sum_{j=1}^{n} y_j^2 = \sum_{j=1}^{n} y_j^2 (1 - R^2) + R^2 \sum_{j=1}^{n} y_j^2$$

ซึ่งจะได้ $$1 = (1 - R^2) + R^2$$

ค่า $(1 - R^2)$ เป็นสัดส่วนของความแปรปรวนของค่าตัวแปรตาม Y ซึ่งไม่ขึ้นกับความแปรปรวนของตัวแปรอิสระทุกตัวในแบบจำลอง (coefficient of nondetermination)

ค่า R^2 คือค่าสัมประสิทธิ์ของการตกลงใจ (coefficient of determination) แสดงสัดส่วนของ y^2 ที่มีความแปรปรวนขึ้นกับความแปรปรวนของตัวแปรอิสระทุกตัวในแบบจำลอง

ค่า R^2 มีค่าอยู่ระหว่าง ๐ ถึง ๑

ถ้าค่า R^2 มีค่าเข้าใกล้ ๑ แสดงว่า ความแปรปรวนของค่าตัวแปรตาม Y ที่เกิดขึ้นเนื่องจากความแปรปรวนของค่า X_i ต่าง ๆ ในแบบจำลอง มีมาก หรือก็คือ แบบจำลองความถดถอยที่ใช้ในการประมาณค่าตัวแปรตาม Y นั้น ให้ค่าประมาณของ Y ได้ใกล้เคียงกับค่าจริงมาก ซึ่งแสดงว่าแบบจำลองที่ใช้ในการประมาณค่าตัวแปรตาม Y นั้น มีความเหมาะสมดีนั่นเอง

ถ้าค่า R^2 มีค่าน้อยมากจนเข้าใกล้ ๐ ก็แสดงว่า สัดส่วนความแปรปรวนของค่าตัวแปรตาม Y ที่ขึ้นอยู่กับความแปรปรวนของค่าตัวแปรอิสระ X_i ต่าง ๆ ในแบบจำลองความถดถอยที่ใช้ มีน้อยมาก ซึ่งย่อมแสดงว่า แบบจำลองความถดถอยที่ใช้ในการประมาณค่าตัวแปรตาม Y นั้น ยังไม่เหมาะสมที่จะนำมาใช้ จะต้องแก้ไข เปลี่ยนแปลงใหม่

ค่าสหสัมพันธ์พาเชียล (Partial Correlation) หรือสหสัมพันธ์เชิงส่วน

ในแบบจำลองความถดถอยที่มีตัวแปรหลายตัว นอกจากเราจะหาค่าสหสัมพันธ์เชิงเดียวระหว่างตัวแปรต่าง ๆ เป็นคู่ ๆ ได้ เช่น r_{ik} ซึ่งก็คือค่าสหสัมพันธ์เชิงเดียว ระหว่างตัวแปรที่ i กับตัวแปรที่ k โดยอาศัยทฤษฎีที่ได้กล่าวมาข้างต้นแล้ว ในกรณีที่ตัวแปรต่าง ๆ ทุกตัว ผันแปรร่วมกัน แบบ การกระจายปกติ (normal distribution) เราก็อาจจะหาค่าสหสัมพันธ์เชิงส่วนของตัวแปรได้ ลองพิจารณาในแบบจำลองความถดถอยที่มีตัวแปรตาม ๑ ตัว และตัวแปรอิสระ ๓ ตัว ซึ่งมีแบบจำลองความถดถอยดังนี้

$$\hat{Y} = b_o + b_1X_1 + b_2X_2 + b_3X_3$$

นอกจากเราจะได้ค่าสหสัมพันธ์เชิงเดียว เช่น r_{YX_1} , r_{YX_2} , $r_{X_1X_2}$ แล้ว เรายังสามารถหาค่า สหสัมพันธ์เชิงส่วนได้ เช่น

$r_{Y1.2}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับตัวแปร X_1 โดยถือว่า X_2 คงที่

$r_{Y2.1}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับตัวแปร X_2 โดยถือว่า X_1 คงที่

$r_{12.3}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร X_1 กับตัวแปร X_2 โดยถือว่า X_3 คงที่

$r_{13.2}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร X_1 กับตัวแปร X_3 โดยถือว่า X_2 คงที่

.....

ค่าสหสัมพันธ์เชิงส่วนต่าง ๆ นั้น จะหาได้โดยอาศัยสูตรต่อไปนี้

$$r_{Y3.12} = \frac{r_{Y3.1} - r_{Y2.1}r_{32.1}}{\left[(1-r^2_{Y2.1})(1-r^2_{32.1})\right]^{\frac{1}{2}}} = \frac{r_{Y3.2} - r_{Y1.2}r_{31.2}}{\left[(1-r^2_{Y1.2})(1-r^2_{31.2})\right]^{\frac{1}{2}}}$$

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\left[(1-r^2_{Y2})(1-r^2_{12})\right]^{\frac{1}{2}}}$$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\left[(1-r^2_{13})(1-r^2_{23})\right]^{\frac{1}{2}}}$$

สำหรับค่าอื่น ๆ ก็คงพิจารณาจากรูปสูตรเดียวกัน

ค่าสหสัมพันธ์เชิงส่วน มีค่าอยู่ระหว่าง -1 ถึง $+1$ และการพิจารณาความหมายของความสัมพันธ์ของตัวแปร จากค่าสหสัมพันธ์เชิงส่วนที่ได้นี้ ก็คงอาศัยหลักการเดียวกับการพิจารณาความหมายของค่าสหสัมพันธ์เชิงเดียว ดังได้กล่าวมาแล้ว

วิธีต่าง ๆ ในการสร้างสมการถดถอย

ในกรณีที่มีตัวแปรตาม และตัวแปรอิสระเพียงตัวเดียว การสร้างแบบจำลองความถดถอยย่อมไม่มีปัญหายุ่งยาก แต่ในการสร้างแบบจำลองความถดถอยในกรณีที่มีตัวแปรอิสระ หลาย ๆ ตัวนั้น ย่อมมีปัญหาเกิดขึ้นว่า เราจะมีวิธีการอย่างไรที่จะสร้างแบบจำลองความถดถอยให้ได้ผลในการประมาณค่าตัวแปรตามได้ดีที่สุด และเป็นวิธีการที่ประหยัด เวลา และ แรงงาน ได้มากที่สุด

วิธีต่าง ๆ ที่นิยมใช้กันในการสร้างสมการถดถอย สำหรับกรณีที่มีตัวแปรอิสระ หลายตัวนั้นมีดังนี้

๑. วิธีใช้ทดสอบแบบจำลองทั้งหมดที่จะสามารถสร้างขึ้นได้ เพื่อหาแบบจำลองที่ดีที่สุด (All Possible Regression)

๒. วิธีใช้ตัวแปรอิสระทุกตัวสร้างแบบจำลองความถดถอยขึ้นมาก่อน แล้วจึงทดสอบนัยสำคัญเพื่อจะตัดตัวแปรอิสระออกไปทีละตัว (ตัดเฉพาะตัวที่ทดสอบแล้วไม่มีนัยสำคัญ) (Backward Selection)

๓. วิธีทดลองเลือกตัวแปรอิสระที่มีค่าสหสัมพันธ์กับตัวแปรตามสูงสุดสร้างแบบจำลองความถดถอยอย่างง่ายขึ้นก่อน แล้วค่อยเพิ่มตัวแปรอิสระตัวที่มีความสำคัญรองลงไปทีละตัวทดสอบดูนัยสำคัญ (Forward Selection) ซึ่งจะนำมาใช้ในกรณีนี้ **(อธิบายหน้า ๒๑)**

๔. วิธีทดสอบทีละตัว ลองเพิ่มเข้าไปและสลับตำแหน่ง (Stepwise Regression)

จากวิธีการต่าง ๆ ดังกล่าวผู้วิจัยพิจารณาเห็นว่า วิธีเลือกตัวแปรอิสระเพิ่มเข้าทีละตัว มีความเหมาะสมกับปัญหานี้มากกว่าวิธีอื่น กล่าวคือวิธีพิจารณาแบบจำลองทั้งหมดที่จะเป็นไปได้นั้น ในขั้นแรกก็ต้องพิจารณาแบบจำลองที่มีตัวแปรอิสระเพียงตัวเดียว ว่าแบบจำลองที่ใช้ตัวแปรอิสระใดจะให้ผลดีที่สุด ซึ่งเมื่อพิจารณาจะเห็นว่าก็คือวิธีการขั้นแรกของวิธีที่จะนำมาใช้นั่นเอง จากที่ได้ทดลองแบบจำลองที่มีตัวแปรอิสระเพียงตัวเดียวแล้ว ขั้นต่อไปก็ต้องพิจารณาแบบจำลองทุกแบบที่จะมีได้ว่าแบบใดดีที่สุด ผู้วิจัยเห็นว่าในขั้นนี้ผลที่ได้ก็คงจะคล้ายคลึงกับวิธีเลือกทีละตัวเพราะเราก็ได้เลือกตัวที่มีความสัมพันธ์กับตัวแปรตามสูงที่สุดอยู่แล้ว ดังนั้นวิธีแรกนี้จึงออกจะยุ่งยาก และผลก็จะไม่แตกต่างกับผลที่ได้จากวิธีเลือกทีละตัวโดยเฉพาะอย่างยิ่งในรูปแบบที่ว่าตัวแปรอิสระตัวใดจะเรียงลำดับก่อนหลังกันอย่างไรในแบบจำลองความถดถอยที่จะนำมาใช้นั้นก็คงจะไม่ต่างกัน

ในวิธีสร้างแบบจำลองที่มีทุกตัวแปรอิสระขึ้นก่อนแล้วตัดออกทีละตัวนั้นนับว่าเป็นวิธีที่ไม่เหมาะอย่างยิ่ง เพราะในการสร้างแบบจำลองรวมนั้นเราจะไม่ได้พิจารณาถึงความสำคัญของตัวแปรอิสระแต่ละตัว ว่าควรจะเรียงลำดับกันอย่างไร แบบจำลองที่ได้จากวิธีนี้จึงไม่ใช่แบบจำลองที่เหมาะสมที่สุด วิธีนี้น่าจะใช้ในกรณีที่มีตัวแปรอิสระมาก ๆ

สำหรับวิธีเลือกเข้าไปทีละตัวและทดลองสลับตำแหน่งนั้น จะเห็นได้ว่าตัวแปรอิสระของเรามีไม่มาก ในขั้นแรกเราก็ได้เลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามสูงที่สุดมาใช้ก่อน ต่อไปก็พิจารณาจากตัวแปรอิสระที่เหลืออยู่ที่มีค่าสหสัมพันธ์เชิงส่วนกับตัวแปรตามสูงที่สุดมาใช้ต่อไป และจากที่ได้พิจารณาเห็นว่า ตัวแปรอิสระที่สัมพันธ์กับตัวแปรตามอยู่ในเกณฑ์สูงนั้น มีเพียงสองตัว ซึ่งก็ได้ทดลองสร้างแบบจำลองของทั้งสองตัวแปรนี้เปรียบเทียบกันก่อนจะนำไปใช้เป็นแบบจำลองจริงแล้ว โดยเลือกตัวแปรที่ให้ผลดีที่สุด ดังนั้นปัญหาในการที่จะสลับที่ของตัวแปรทั้งสองน่าจะไม่มี และสำหรับตัวแปรที่ ๓ และที่ ๔ ก็มีความสัมพันธ์กับตัวแปรตามแตกต่างกันมาก โดยเฉพาะตัวแปรอิสระตัวที่ ๔ นั้น มีความสัมพันธ์กับตัวแปรตามอยู่ในเกณฑ์ต่ำมาก และเมื่อนำไปสร้างแบบจำลองความถดถอยทำการทดสอบแล้วก็ไม่มีนัยสำคัญทางสถิติ ดังนั้นปัญหาในการสลับตำแหน่งจึงหมดไป

จากการพิจารณาถึงผลของการเลือกใช้วิธีต่าง ๆ ในการสร้างสมการถดถอยแล้ว ผู้วิจัยจึงได้ตกลงใจใช้วิธีเลือกตัวแปรเพิ่มเข้าไปทีละตัว เป็นวิธีวิเคราะห์ความสัมพันธ์ในปัญหานี้ ดังจะได้อธิบายวิธีที่จะนำมาใช้นี้ในหัวข้อต่อไป

วิธี Forward Selection Procedure

เป็นวิธีในการสร้างสมการถดถอยสำหรับปัญหาที่มีตัวแปรอิสระหลาย ๆ ตัว โดยเลือกสร้างสมการถดถอยเชิงเดียว (simple regression)ระหว่างตัวแปรตาม กับ ตัวแปรอิสระตัวที่มีค่าสหสัมพันธ์เชิงเดียวกับตัวแปรตามนั้นสูงสุด หรือก็คือเลือกตัวแปรอิสระที่เห็นว่ามีความสำคัญกับตัวแปรตามมากที่สุดนั่นเองมาพิจารณาก่อน แล้วทำการทดสอบแบบจำลองที่ได้ ถ้าพบว่ามีนัยสำคัญก็ดำเนินการต่อไป โดยพิจารณาเลือกตัวแปรอิสระที่เหลืออยู่ ซึ่งมีค่าสหสัมพันธ์เชิงส่วนกับตัวแปรตามนั้นสูงสุด (โดยไม่พิจารณาถึงตัวแปรอิสระตัวที่ได้ใช้ในแบบจำลองไปแล้ว) เพิ่มเข้าไปเป็นตัวแปรอิสระอีกตัวหนึ่ง และสร้างแบบจำลองความถดถอยใหม่ขึ้น ทำการทดสอบนัยสำคัญต่อไปว่าแบบจำลองความถดถอยใหม่ที่สร้างขึ้นโดยเพิ่มตัวแปรอิสระตัวใหม่เข้าไปนี้ จะมีความสำคัญเป็นนัยสำคัญในการประมาณค่าตัวแปรตาม Y หรือไม่ และแบบจำลองใหม่นี้ให้ผลดีขึ้นกว่าแบบจำลองเดิมหรือไม่ จากนั้นก็พิจารณาเลือกตัวแปรอิสระตัวต่อไป สร้างแบบจำลองความถดถอยขึ้นใหม่และทดสอบ ตามวิธีที่กล่าวมาแล้ว ตัวแปรอิสระตัวใดที่ทดสอบในแบบจำลองแล้วไม่มีนัยสำคัญก็จะตัดทิ้งออกไปไม่นำมาใส่ในแบบจำลอง

วิธีนี้จะจบลงเมื่อพบว่า ตัวแปรอิสระที่เหลืออยู่เมื่อนำไปเพิ่มสร้างแบบจำลองใหม่สำหรับใช้ในการประมาณค่าตัวแปรตาม Y แล้ว แบบจำลองที่ได้ใหม่ไม่มีนัยสำคัญ หรือก็คือตัวแปรอิสระตัวที่เพิ่มเข้าไปใหม่นั้น ไม่มีความสัมพันธ์เป็นนัยสำคัญกับตัวแปรตาม Y หรืออาจจะพิจารณาจากค่า สัมประสิทธิ์แห่งการตกลงใจ R^2 ของแบบจำลองที่ทดสอบได้นัยสำคัญนั้นว่ามีค่าสูงพอหรือไม่