บทที่ 3
ทฤษฎีที่เกี่ยวข้องกับการวิจัย

3.1 บทนำ

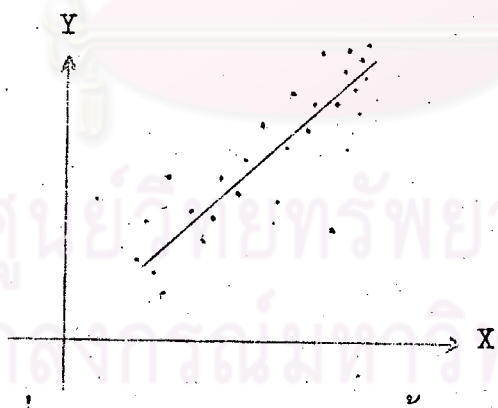
ในปัญหาทางสถิติหลายชนิดที่เกิดขึ้น เรามักจะสนใจตัวแปรตัวหนึ่งซึ่งเป็นตัวแปรที่ขึ้นอยู่กับตัวแปรอีกตัวหนึ่งหรือหลายตัว ตัวอย่างเช่น ในปัญหาทางการเกษตร เราต้องการตรวจสอบว่า ผลผลิตที่ได้จากการปลูกข้าวในนาทดลองแห่งหนึ่งจะขึ้นอยู่กับปริมาณน้ำ หรือปริมาณปุ๋ยที่ใช้ ในทำนองเดียวกัน ปัญหาที่เกี่ยวกับการศึกษาเราอาจต้องการที่จะหาความสัมพันธ์ระหว่างคะแนนสอบคัดเลือกของนักเรียนที่สอบเข้ามหาวิทยาลัยได้กับผลการเรียนในชั้นปีที่ 1 ว่ามีความสัมพันธ์กันแบบไหน หรือในปัญหาทางเศรษฐศาสตร์ นักเศรษฐศาสตร์อาจต้องการที่จะศึกษาถึงความแปรปรวนของความต้องการเครื่องอุปโภคบริโภคในช่วงระยะปีหนึ่ง ๆ ว่าเป็นอย่างไร เป็นต้น

ในแต่ละตัวอย่างดังที่ยกมานี้ เราจะพิจารณาตัวแปรตัวหนึ่งซึ่งเราเรียกว่าตัวแปรตาม (dependent variable) ว่าเป็นฟังก์ชันของตัวแปรตัวอื่น ซึ่งเราจะเรียกว่าตัวแปรอิสระ (independent variable) ถ้าเราสามารถหาความสัมพันธ์ระหว่างตัวแปร 2 ชนิดนี้ได้แล้ว จะทำให้สามารถทำนายค่าของตัวแปรตาม เมื่อทราบค่าของตัวแปรอิสระได้ อย่างเช่น ในตัวอย่างที่เกี่ยวกับการศึกษา ถ้าเราสามารถหาความสัมพันธ์ระหว่างคะแนนสอบคัดเลือกเข้ามหาวิทยาลัย กับผลการเรียนในชั้นปีที่ 1 ได้แล้ว เราจะสามารถทำนายได้ว่านักศึกษาคนหนึ่ง ๆ จะมีผลการเรียนในชั้นปีที่ 1 เป็นอย่างไร ถ้าเราทราบคะแนนสอบคัดเลือกเข้ามหาวิทยาลัยของเขา.

ปัญหาในทำนองนี้ เรียกว่าเป็นปัญหาที่เกี่ยวกับความสัมพันธ์เชิงฟังก์ชัน (Functional relationship) ปัญหาเหล่านี้จะซับซ้อนขึ้น ถ้าใช้ตัวแปรอิสระหลาย ๆ

ตัวร่วมกันทำนายลักษณะของตัวแปรตาม จากตัวอย่างที่เกี่ยวกับการศึกษา ถ้าผลการเรียนในชั้นปีที่ 1 ในมหาวิทยาลัยขึ้นอยู่กับคะแนนสอบคัดเลือกเข้ามหาวิทยาลัย, และคะแนนสอบไล่ชั้นประโยคมัธยมศึกษาตอนปลายแล้ว เราจะใช้ตัวแปรอิสระ 2 ตัว คือ ผลการสอบคัดเลือกและคะแนนสอบไล่ชั้นประโยคมัธยมศึกษาตอนปลาย ร่วมกันทำนายผลการเรียนในชั้นปีที่ 1 ซึ่งอาจจะทำให้สามารถทำนายลักษณะของตัวแปรตาม ได้ดีกว่าที่จะใช้ตัวทำนายเพียงตัวเดียว

การศึกษาความสัมพันธ์ของตัวแปร 2 ชนิด อาจจะแสดงความสัมพันธ์ขั้นต้นได้โดยการนำค่าของตัวแปรทั้ง 2 ไปเขียนกราฟ ซึ่งจะได้จุดการกระจายทั่วไป เรียกว่า แผนภาพกระจาย (Scatter diagram) แล้วเขียนเส้นตรงหรือเส้นโค้งแสดงแนวโน้มโดยประมาณของความสัมพันธ์ เส้นที่ได้เรียกว่า เส้นถดถอย (Regression line or curve) และสมการของเส้นเรียกว่า สมการถดถอย (Regression Equation) เช่นอาจได้ดังแผนภาพการกระจายที่แสดงไว้ข้างล่างนี้



รูปที่ 1 แผนภาพการกระจายของข้อมูล

การถดถอย (Regression) จึงเป็นการศึกษาดูว่า ตัวแปร 2 ตัว มีความสัมพันธ์กันหรือไม่ ถ้าสัมพันธ์กันแล้วสัมพันธ์กันในลักษณะใด

3.2 การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)

3.2.1 แบบการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Model)

ในกรณีที่เราต้องการจะศึกษาว่าตัวแปร 2 ตัว คือ ตัวแปรตาม (Y) มีความสัมพันธ์กับตัวแปรอิสระ 1 ตัว (X) ในเชิงเส้นตรงหรือไม่นั้น เป็นการศึกษาถึงการถดถอยเชิงเส้นอย่างง่าย

ถ้า Y และ X มีความสัมพันธ์กันเชิงเส้นตรง เราจะเขียนความสัมพันธ์ระหว่างตัวแปรทั้งสอง ในรูป

$$Y = \beta_0 + \beta_1 X + \epsilon$$

โดยที่ β_0 คือค่าที่เส้นถดถอยตัดแกน Y (Y intercept)

β_1 คือสัมประสิทธิ์แห่งการถดถอยของประชากร (Population regression coefficient) ซึ่งเป็นพารามิเตอร์ที่ยังไม่ทราบค่า

ϵ คือความคลาดเคลื่อน (error) ซึ่งเป็นตัวแปรสุ่มที่ยังไม่ทราบค่าเป็นจำนวนที่ค่าจริงของ Y ต่างไปจากค่าประมาณของ Y

และ $E(\epsilon_i) = 0, \quad i = 1, 2, \ldots, n$

$E(\epsilon_i, \epsilon_j) = \sigma^2, \quad i = j = 1, 2, \ldots, n$

$= 0, \quad i \neq j$

3.3.2 การประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุด

ในตัวแบบ $Y = \beta_0 + \beta_1 X + \epsilon$

ถ้าให้ b_0 และ b_1 เป็นค่าประมาณของ B_0 และ B_1

จะได้ $\hat{Y} = b_0 + b_1 X$ เป็นสมการถดถอย

ในการประมาณค่า β_o, β_1 ด้วย b_o, b_1 นั้น เราใช้วิธีกำลังสองน้อยที่สุด (Method of least squares) โดยอาศัยข้อมูลที่รวบรวมมาได้

สมมติว่า ข้อมูลที่รวบรวมได้คือ $(X_1, Y_1,), (X_2, Y_2),..., (X_n, Y_n)$ โดยที่

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i , \quad i = 1, 2,.., n$$

ผลบวกของกำลังสองของผลต่างจากเส้นที่แท้จริง (sum of squares of deviations from the true line) คือ

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2 \quad \ldots\ldots 1$$

ตามวิธีกำลังสองน้อยที่สุดนั้นเราจะพยายามทำให้ค่า S น้อยที่สุดจากการดิฟเฟอเรนชิเอท (Differentiate) สมการ 1 เทียบกับ β_o ครั้งหนึ่ง แล้วเทียบกับ β_1 อีกครั้งหนึ่ง ได้

$$\frac{\partial S}{\partial \beta_o} = -2 \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i) \quad \ldots\ldots 2$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_o - \beta_1 X_i) \quad \ldots\ldots 3$$

เทียบสมการ 2 และ 3 ให้เท่ากับ 0 แล้วแทนค่า β_o และ β_1 ด้วย b_o และ b_1 ได้สมการ

$$\sum_{i=1}^{n} (Y_i - b_o - b_1 X_i) = 0 \quad \ldots\ldots 4$$

$$\sum_{i=1}^{n} X_i (Y_i - b_o - b_1 X_i) = 0 \quad \ldots\ldots 5$$

จาก 4 และ 5 ได้

$$\sum_{i=1}^{n} Y_i - nb_o - b_1 \sum_{i=1}^{n} X_i = 0 \qquad \ldots\ldots 6$$

$$\sum_{i=1}^{n} X_i Y_i - b_o \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0 \qquad \ldots\ldots 7$$

หรือ

$$b_o n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad \ldots\ldots 8$$

$$b_o \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \qquad \ldots\ldots 9$$

เรียกสมการที่ได้เหล่านี้ว่า <u>สมการปกติ</u> (normal equations)

จาก 8 และ 9 หาค่า b_1 และ b_o ออกมาได้ดังนี้

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right] / n}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 / n}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad \ldots\ldots 10$$

และ $b_o = \bar{Y} - b_1 \bar{X}$ 11

∴ จากสมการ $\hat{Y}_i = b_o + b_1 X_i$

ได้ $\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X})$ 12

เมื่อพิจารณาความคลาดเคลื่อน ϵ_i ซึ่งคือ $Y_i - \hat{Y}$ (ความคลาดเคลื่อนที่เกิดจากค่าประมาณต่างไปจากค่าจริง)

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

จะเห็นว่า ความคลาดเคลื่อน เกิดจากค่าจริงเบี่ยงเบนไปจากค่าเฉลี่ยและอีกส่วนหนึ่งเกิดจากค่าประมาณเบี่ยงเบนไปจากค่าเฉลี่ย

จาก $\epsilon_i = (Y_i - \hat{Y}_i) = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})$

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[(Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})\right]^2$$

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

พิจารณาเทอม $\sum_{i=1}^{n} (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})$ 13

แทนค่า $\hat{Y}_i$ ด้วยค่าทางขวามือในสมการ 12

$$\therefore \sum_{i=1}^{n} (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (Y_i - \bar{Y})\left[\bar{Y} + b_1(X_i - \bar{X}) - \bar{Y}\right]$$
$$= b_1 \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})$$

[จากสมการ 10 $b_1 = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n} (X_i - \bar{X})^2$

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

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

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

[จากสมการ 12 จะได้ $\hat{Y}_i - \bar{Y} = b_1 (X_i - \bar{X})$

$$(\hat{Y}_i - \bar{Y})^2 = b_1^2 (X_i - \bar{X})^2$$

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} b_1^2 (X_i - \bar{X})^2 \quad]$$

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

ดังนั้น จากสมการ 13 จะได้

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

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

ดังนั้น $$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \quad \ldots\ldots 14$$

จาก 14 จะเรียก $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ ว่า ผลบวกของกำลังสองของทั้งหมด (Total sum of squares) หรือ ผลบวกของกำลังสองเกี่ยวกับค่าเฉลี่ย (Sum of squares about mean)

$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ว่า ผลบวกของกำลังสองเกี่ยวกับการถดถอย (Sum of squares about regression) หรือ ผลบวกของกำลังสองของความคลาดเคลื่อน (Error sum of squares)

$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ ว่า ผลบวกของกำลังสองเนื่องมาจากการถดถอย (Sum of squares due to regression)

ถ้าให้ $R^2 = \dfrac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} =$ ผลบวกของกำลังสองเนื่องมาจากการถดถอย / ผลบวกของกำลังสองของทั้งหมด

เนื่องจากการถดถอยของเราจะดีถ้า ค่าประมาณต่างจากค่าจริงน้อย

นั่นคือ ถ้า $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ เล็ก

ถ้าค่า $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ยิ่งเล็ก แสดงว่าค่า R^2 จะเข้าใกล้ 1 มากขึ้น

ฉะนั้น ถ้า R^2 เข้าใกล้ 1 แล้ว เราจะได้ตัวแบบ (model) ที่ดี ค่า R^2 จะอยู่ระหว่าง $0 \longrightarrow 1$

3.2.3 <u>การทดสอบสมมุติฐาน</u> (Tests of Hypotheses)

<u>การทดสอบสมมุติฐานโดยใช้การทดสอบ F</u> (F - test)

ถ้าเรามีตัวแบบ

$$Y = \beta_0 + \beta_1 X + \epsilon$$

จะทดสอบสมมุติฐานว่า ตัวแบบที่ใช้เหมาะสมหรือไม่ คือ จะทดสอบว่ามีการถดถอยหรือไม่นั้น จะตั้งสมมุติฐานว่า

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$$

โดยมีข้อสมมุติว่า ϵ_i มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 0 และมีความแปรปรวนเป็น σ^2

สร้างตารางวิเคราะห์ความแปรปรวน (Table of Analysis of variance) เพื่อการทดสอบสมมุติฐานดังกล่าว จะเรียกย่อ ๆ ว่า ตาราง ANOVA ดังนี้

ตารางที่ 3.1

ANOVA

Source of variations	d.f.	Sum of squares	Mean squares	F
Due to regression	1	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = SSR$	SSR/1 = MSR	$\frac{MSR}{MSE}$
About regression (residual)	n-2	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = SSE$	SSE/n-2 = MSE	
Total	n-1	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = SST$	SST/n-1	

สูตรสำหรับการคำนวณ

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

$$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = b_i\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

$$= b_i\left[\sum_{i=1}^{n}X_iY_i - \frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}\right]$$

หรือ $$= b_i^2\sum_{i=1}^{n}(X_i-\bar{X})^2$$

สำหรับ $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ หาได้จาก $\sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$

ค่า F ที่ได้จากการคำนวณ MSR/MSE นำมาเปรียบเทียบกับค่า F จากตารางที่ขั้นแห่งความเป็นอิสระ (degree of freedom) (1, n-2)

ถ้า $F_{\text{คำนวณ}} > F_{\text{ตาราง}}$ เราจะไม่ยอมรับสมมุติฐานที่ว่า $\beta_1 = 0$

นั่นคือ $\beta_1 \neq 0$ แสดงว่า มีการถดถอย

จากตารางวิเคราะห์ความแปรปรวน ค่า MSE (Mean Squares of Error) ซึ่งเท่ากับ $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ หารด้วยขั้นแห่งความเป็นอิสระ จะเป็นค่าความแปรปรวนของ Y เมื่อคำนึงถึง X ด้วย แทนด้วยสัญญลักษณ์ $s^2_{Y.X} = 6^2$

และจะได้ว่า $s^2_Y = \sum_{i=1}^{n}\frac{(Y_i - \bar{Y})^2}{n-1}$

<u>การทดสอบสมมุติฐานโดยใช้การทดสอบ t</u> (t - test)

ถ้ามีตัวแบบเป็น $Y_i = \beta_o + \beta_1 X_i + \epsilon_i$

โดยที่ ϵ_i มีการแจกแจงแบบปกติมีค่าเฉลี่ยเป็น 0 มีความแปรปรวน 6^2

b_1 จะมีการกระจายเป็นแบบปกติ (normal) มีค่าเฉลี่ยเป็น β_1 และความแปรปรวนเป็น $6^2/\sum_{i=1}^{n}(X_i - \bar{X})^2$

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

$$V(b_1) = \sum_{i=1}^{n}\frac{6^2}{(X_i - \bar{X})^2}$$

ซึ่งประมาณด้วย $s^2_{b_1} = s^2_{Y.X}\Big/\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$(b_1 - \beta_1)\Big/\sqrt{V(b_1)} \sim N(0,1)$$

ซึ่งประมาณด้วย $$\left((b_1 - \beta_1)\Big/\sqrt{\frac{s^2_{Y.X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right) \sim t_{(n-2)}$$

ถ้าเราตั้งสมมุติฐานว่า

$$H_o : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

จาก $$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$= \frac{b_1}{s_{b_1}}$$

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

เราจะยอมรับสมมุติฐานนี้ ถ้า $t_{\alpha/2} < t < t_{1-\alpha/2}$

โดยที่ α = ระดับที่มีนัยสำคัญ

3.3 <u>แบบเชิงเส้นทั่วไป</u> (The General Linear Model) <u>หรือการถดถอยพหุคูณ</u> (Multiple Regression)

จากปัญหาทางสถิติโดยทั่ว ๆ ไป ที่จำเป็นต้องใช้การถดถอย (Regression) ในการวิเคราะห์ปัญหานั้น การศึกษาถึงการถดถอยเชิงเส้นอย่างง่าย (Simple linear regression) ซึ่งเป็นการศึกษาถึงความสัมพันธ์ของตัวแปรเพียง 2 ตัว คือ ตัวแปรตาม และตัวแปรอิสระอีกเพียงตัวเดียว ดังได้กล่าวในข้อ 3.2 นั้น มักไม่เป็นการเพียงพอ ทั้งนี้เพราะการที่จะประมาณค่าของ Y ให้ได้ใกล้เคียงที่สุดนั้น เรามักจะพิจารณาถึงตัวแปรอิสระหลาย ๆ ตัว ที่มีอิทธิพลต่อ Y การศึกษาถึงตัวแปรอิสระที่มากกว่า 1 ตัวขึ้นไป ที่มีอิทธิพลหรือมีความสำคัญต่อ Y นี้ เราจะใช้การถดถอยชนิดที่เรียกว่า การถดถอยพหุคูณ (Multiple regression) หาสมการที่จะใช้ทำนายค่า Y เมื่อมีค่า X มากกว่า 1 ตัว ดังเช่นในการวิเคราะห์ปัญหาของการวิจัยครั้งนี้ ก็จะใช้วิธีการถดถอยพหุคูณ

3.3.1 <u>แบบการถดถอยพหุคูณ</u> (Multiple Regression Model)

สมมุติว่า มีความสัมพันธ์เชิงเส้นระหว่างตัวแปร Y ซึ่งเป็นตัวแปรตาม กับตัวแปร X ซึ่งเป็นตัวแปรอิสระ k ตัว คือ $X_1, X_2, \ldots, X_k$ และความคลาดเคลื่อน (disturbance term) ϵ แล้ว เราสามารถเขียนความสัมพันธ์เหล่านี้ได้ในรูป

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon \quad \ldots\ldots 15$$

ตัวที่ยังไม่ทราบค่าคือ β_i และพารามิเตอร์ที่ได้จากการแจกแจง ϵ จะประมาณค่าเหล่านี้ออกมา โดยอาศัยข้อมูลที่รวบรวมได้คือ $(Y_i, X_{1_i}, X_{2_i}, \ldots, X_{k_i})$, $i = 1, 2, \ldots\ldots, n$ ทำให้เขียน 15 ได้ในรูป

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \qquad \ldots\ldots 16$$

โดยที่

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} \cdots & X_{k1} \\ 1 & X_{12} & X_{22} \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \cdots & X_{kn} \end{bmatrix}$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

จากสมการ 16 จะมีข้อสมมุติ (Assumptions) ว่า

(1) $E(\underline{\epsilon}) = \underline{0}$

(2) $E(\underline{\epsilon}\underline{\epsilon}') = \sigma^2 I$

(3) X เป็นกลุ่มของจำนวนคงที่ (fixed numbers)

(4) X มี rank $k + 1 < n$

จากข้อ (1) $E(\underline{\epsilon}) = \underline{0}$ แสดงว่า $E(\epsilon_i) = 0$ สำหรับทุกค่า i

จากข้อ (2) $E(\underline{\epsilon}\underline{\epsilon}') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \vdots & & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{bmatrix}$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

แสดงว่า

$E(\epsilon_i^2) = \sigma^2$ สำหรับทุกค่า i นั่นคือ ϵ_i มีความแปรปรวนคงที่

$E(\epsilon_i\epsilon_j) = 0$ เมื่อ $i \neq j$ แสดงว่าแต่ละคู่ของ ϵ_i ไม่เกี่ยวข้องกัน (pairwise uncorrelated)

จากข้อ (3) $\underline{X}$ เป็นกลุ่มของจำนวนคงที่ หมายความว่า ในการประมาณค่า Y นั้น เราจะต้องทราบค่า X ก่อน นั่นคือ X_i แต่ละตัวเป็นจำนวนคงที่

จากข้อ (4) $\underline{X}$ มี rank $k + 1 < n$ หมายความว่า จำนวนค่าสังเกต (n) จะมากกว่าจำนวนพารามิเตอร์ $(k + 1)$ที่จะต้องประมาณค่าออกมา

3.3.2 <u>การประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุด</u>

ให้ $\hat{\underline{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$

จะได้ $\underline{Y} = \underline{X}\hat{\underline{\beta}} + \underline{e}$ 17

โดยที่ $\underline{e}$ เป็นเวกเตอร์สดมภ์ (column vector) ของ ส่วนที่เหลือ n ค่า $(\underline{Y} - \underline{X}\hat{\underline{\beta}})$

จาก 17 ได้ผลบวกของส่วนที่เหลือยกกำลังสองเป็น

$$\sum_{i=1}^{n} e_i^2 = \underline{e}'\underline{e}$$

$$= (\underline{Y} - \underline{X}\hat{\underline{\beta}})'(\underline{Y} - \underline{X}\hat{\underline{\beta}})$$

$$= \underline{Y}'\underline{Y} - 2\hat{\underline{\beta}}'\underline{X}'\underline{Y} + \hat{\underline{\beta}}'\underline{X}'\underline{X}\hat{\underline{\beta}}$$

เพื่อที่จะหาค่า $\hat{\underline{\beta}}$ ออกมา จะใช้วิธีที่ทำให้ค่า $\sum_{i=1}^{n} e_i^2$ มีค่าน้อยที่สุด (minimize) ดังนี้

$$\frac{\partial(\underline{e}'\underline{e})}{\partial\hat{\underline{\beta}}} = -2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X}\hat{\underline{\beta}}$$

แล้วเทียบให้เท่ากับ 0 จะได้

$$\underline{X}'\underline{X}\hat{\underline{\beta}} = \underline{X}'\underline{Y}$$

$$\therefore \quad \hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} \qquad \ldots\ldots 18$$

ถ้า $(\underline{X}'\underline{X})^{-1}$ เป็นแมตริกซ์ไม่เอกเทศ (nonsingular matrix) นั่นคือจะหาค่าประมาณของ $\underline{\beta}$ หรือ $\hat{\underline{\beta}}$ ได้จากสูตรที่ 18 นี้

3.3.3 การทดสอบสมมุติฐาน (Tests of Hypotheses)

ในเรื่องที่เกี่ยวกับการถดถอยนั้น เรามักจะพิจารณาว่าจะเป็นการตุ้มค่าหรือไม่ ถ้าจะเพิ่ม X_i เข้าไปในตัวแบบอีกตัวหนึ่ง ในการพิจารณาเรื่องดังกล่าว เรามักจะพิจารณาที่ผลบวกของกำลังสอง อันเนื่องมาจากการถดถอย (Regression sum of squares) ของเทอมที่เราคิดว่าควรจะอยู่ในตัวแบบหรือไม่ โดยที่ค่าเฉลี่ยกำลังสอง (mean squares) ที่ได้จากผลบวกของกำลังสองเกี่ยวกับการถดถอยนี้ จะถูกนำมาเปรียบเทียบกับ s^2 ซึ่งเป็นค่าประมาณของ σ^2 เพื่อที่จะดูว่าค่าเฉลี่ยกำลังสองใหญ่กว่า s^2 อย่างมีนัยสำคัญหรือไม่ ถ้ามีนัยสำคัญ เราก็จะรวมเทอมที่กำลังพิจารณาอยู่นี้เข้าไปในตัวแบบ แต่ถ้าไม่มีนัยสำคัญ เราก็จะตัดออก

นั่นคือ สมมุติว่า มีฟังก์ชัน $Z_1, Z_2, \ldots, Z_p$ ซึ่งเป็นฟังก์ชันที่ทราบค่า (known function) ของตัวแปร $X_1, X_2, \ldots$ และตัวแปร $X_1, X_2, \ldots$ และ Y เป็นตัวแปรที่สามารถหาค่าได้

พิจารณาตัวแบบทั้ง 2 นี้

$$(1)\quad Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \epsilon$$

สมมุติว่าได้ค่าประมาณกำลังสองน้อยที่สุดของพารามิเตอร์เป็น $b_0(1), b_1(1), b_2(1), \ldots, b_p(1)$ และได้ผลบวกกำลังสองเป็น $SS(b_0(1), b_1(1), b_2(1), \ldots, b_p(1)) = S_1$ โดยที่ s^2 เป็นค่าประมาณของ σ^2

$$(2)\quad Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_q Z_q + \epsilon$$

โดยที่ Z_i ในตัวแบบที่ 2 มีฟังก์ชันเดียวกับ Z_i ที่ตรงกันในตัวแบบที่ 1

สมมุติว่าได้ค่าประมาณกำลังสองน้อยที่สุดของพารามิเตอร์เป็น $b_o(2), b_1(2), b_2(2), \ldots, b_q(2)$ และได้ผลบวกกำลังสองเป็น

$$SS(b_o(2), b_1(2), b_2(2), \ldots, b_q(2)) = S_2$$

จะเรียก $S_1 - S_2$ เป็นผลบวกกำลังสองชนิดพิเศษ (extra sum of squares) อันเนื่องมาจากการรวมเทอม $\beta_{q+1} Z_{q+1} + \ldots + \beta_p Z_p$ เข้าไปในตัวแบบที่ 1 โดยที่ S_1 มีขั้นแห่งความเป็นอิสระ $p + 1$ และ S_2 มีขั้นแห่งความเป็นอิสระ $q + 1$ ดังนั้น $S_1 - S_2$ จะมีขั้นแห่งความเป็นอิสระ $p - q$

และถ้า $\beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$

แล้วจะได้ว่า $E\{(S_1 - S_2)/(p - q)\} = \sigma^2$

และถ้าความคลาดเคลื่อน (errors) มีการแจกแจงเป็นปกติแล้ว $(S_1 - S_2)/(p - q)$ จะมีการแจกแจงเป็น $\sigma^2 \chi^2_{p-q}$ และเป็นอิสระกับ s^2

นั่นคือ เราจะเปรียบเทียบ $(S_1 - S_2)/(p - q)$ กับ s^2 โดยใช้การทดสอบ $F(p - q, \nu)$ โดยที่ ν เป็นขั้นแห่งความเป็นอิสระของ s^2 เพื่อที่จะทดสอบสมมุติฐานที่ว่า

$$H_o : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$$

เราอาจจะเขียน $S_1 - S_2$ ในรูป $SS(b_{q+1}, \ldots, b_p | b_o, b_1, \ldots, b_q)$ ก็ได้ ซึ่งจะอ่านว่า ผลบวกของกำลังสองของ $b_{q+1}, \ldots, b_p$ เมื่อมี $b_o, b_1, \ldots, b_q$ อยู่แล้ว (Sum of squares of $b_{q+1}, \ldots, b_p$ Given $b_o, b_1, \ldots, b_q$)

โดยใช้หลักของผลบวกของกำลังสองชนิดพิเศษนี้ ในแบบการถดถอยอื่น ๆ เราอาจจะหา $SS(b_o)$, $SS(b_1/b_o)$, $SS(b_2/b_o, b_1)$, ..., $SS(b_p/b_o, b_1, ..., b_{p-1})$ ได้ในทำนองเดียวกัน เพื่อประโยชน์ที่จะใช้ในการทดสอบสมมุติฐานตามจุดประสงค์

โดยอาศัยหลักของผลบวกของกำลังสองชนิดพิเศษ ถ้าเราต้องการที่จะหา $SS(b_i | b_o, b_1, ..., b_{i-1}, b_{i+1}, ..., b_k)$ $i = 1, 2, ..., k$ ซึ่งผลบวกของกำลังสองนี้จะมีขั้นแห่งความเป็นอิสระ 1 นั้น หมายความว่าเราจะพิจารณาว่าถ้าจะเพิ่มเทอม β_i เข้าไปในตัวแบบ หรืออีกนัยหนึ่งก็คือ พิจารณาว่า β_i เป็นเทอมที่เพิ่งถูกใส่เข้าไปในตัวแบบเป็นตัวสุดท้าย ค่าเฉลี่ยกำลังสอง (mean squares) ของ β_i จะเท่ากับผลบวกของกำลังสอง เนื่องจากขั้นแห่งความอิสระเป็น 1 เมื่อนำมาเปรียบเทียบกับ s^2 โดยใช้การทดสอบ F (F - test) แล้วในกรณีนี้ เราจะเรียกว่าเป็นการทดสอบ F เพียงบางส่วน (Partial F - test) สำหรับ β_i

จะเห็นว่าการทดสอบ F เพียงบางส่วนมีประโยชน์มากสำหรับนำไปใช้เป็นมาตรฐานในการพิจารณาเกี่ยวกับการเพิ่มหรือตัดตัวแปรออกจากตัวแบบ ตัวแปรอิสระบางตัวอย่างเช่น X_q อาจจะมีอิทธิพลมาก ถ้าสมการถดถอยนั้นประกอบด้วยตัวแปร X_q เพียงตัวเดียว แต่ถ้า X_q เป็นตัวแปรที่ถูกดึงเข้ามาในตัวแบบที่หลังตัวแปรตัวอื่น ๆ X_q อาจจะมีอิทธิพลต่อตัวแปรตัวอื่นได้ ถ้าสหสัมพันธ์ระหว่าง X_q และตัวแปรที่ถูกดึงเข้ามาในตัวแบบก่อนนั้นสูง

ในการใช้การทดสอบ F เพียงบางส่วน จะใช้โดยพิจารณาเหมือนว่าตัวแปรที่นำมาทดสอบนั้น ถูกดึงเข้ามาในตัวแบบเป็นตัวสุดท้าย เพื่อที่จะดูว่ามีผลอันเกิดจากความสัมพันธ์ระหว่างตัวแปรอิสระเองหรือไม่ การใช้การทดสอบ F เพียงบางส่วนจะช่วยตัดสินใจได้ว่าจะเลือกตัวแปรตัวไหนในกรณีที่จำเป็นต้องมีการตัดสินใจเกี่ยวกับเรื่องนี้

สมมุติว่าเรามีตัวแบบ $Y = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p + \epsilon$

ในการทดสอบสมมุติฐานนั้น สมมุติฐานที่น่าสนใจมีหลายชนิด เช่น

(1) $H_0 : \beta_1 = \beta_1^*, \beta_2 = \beta_2^*, \ldots\ldots, \beta_p = \beta_p^*$

(2) $H_0 : \beta_1 = \beta_2 = \ldots\ldots = \beta_p = 0$

(3) $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots\ldots = \beta_p = 0 , q < p$

(4) $H_0 : \beta_k = 0 , 1 \leq k \leq p$

<u>แบบที่ 1</u> $H_0 : \beta_1 = \beta_1^*, \beta_2 = \beta_2^*, \ldots\ldots, \beta_p = \beta_p^*$

สร้างตารางวิเคราะห์ความแปรปรวน

<u>ตารางที่ 3.2</u>

<u>ANOVA</u>

Source of variation	d.f.	Sum of squares	Mean Squares	F
Due to regression	p	$Q_2 = (Y-X\beta^*)'XSX'(Y-X\beta^*)$	$Q_2/p = MSR$	$\frac{MSR}{MSE}$
Residual	n-p-1	$Q_1 = Y'(I - XSX')Y$	$Q_1/(n-p-1) = MSE$	
Total	n-1	$Q = (Y-X\beta^*)'(Y-X\beta^*)$		

$S = X'X$, Y และ X ใช้ deviation form

นำค่า F ที่ได้จากการคำนวณมาเปรียบเทียบกับค่า F จากตารางที่ขั้นแห่งความเป็นอิสระ $(p, n - p - 1)$

ถ้า $F_{\text{คำนวณ}} > F_{\text{ตาราง}}$ เราจะไม่ยอมรับสมมุติฐาน

แบบที่ 2 $H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

สร้างตารางวิเคราะห์ความแปรปรวน

ตารางที่ 3.3

ANOVA

Source of variation	d.f.	Sum of squares	Mean Squares	F
Due to regression	p	$\hat{\beta}' X' Y = SSR$	$SSR/_{p} = MSR$	$\frac{MSR}{MSE}$
Residual	n-p-1	$Y'Y - \hat{\beta}'X'Y = SSE$	$SSE/_{n-p-1} = MSE=s^2$	
Total	n-1	$Y'Y$		

นำค่า F ที่ได้จากการคำนวณ มาเปรียบเทียบกับค่า F จากตารางที่ขั้นแห่งความเป็นอิสระ $(p, n - p - 1)$

ถ้า $F_{\text{คำนวณ}} > F_{\text{ตาราง}}$ เราจะไม่ยอมรับสมมุติฐานที่ว่า $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ นั่นคือ จะมี β_i อย่างน้อย 1 ตัวที่ไม่เท่ากับ 0

แบบที่ 3 $H_o : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, $q < p$

สร้างตารางวิเคราะห์ความแปรปรวน

ตารางที่ 3.4

ANOVA

Source of variation	d.f.	Sum of squares	Mean Squares	F
All p variables	p	$\hat{\beta}' X' Y = SSR$		
First q variables	q	$\hat{\beta}^{*}{}' X_1' Y = SSR_1$		
Difference	p-q	$SSR - SSR_1$	$MSR_1 = \frac{SSR-SSR_1}{p-q}$	$\frac{MSR_1}{MSE}$
Residual	n-p-1	$SSE = YY - SSR$	$MSE = SSE/_{n-p-1} = s^2$	
Total	n-1	YY		

นำค่า F ที่ได้จากการคำนวณมาเปรียบเทียบกับค่า F ที่ได้จากตาราง ที่ขั้นแห่งความเป็นอิสระ $(p - q, n - p - 1)$ วิธีการเช่นเดียวกับแบบที่กล่าวข้างต้น

ในการคำนวณหาค่า $\hat{\beta}^{*}{}' X_1' Y$ หรือ SSR_1 นั้น จะหาได้จาก

ตัวแบบ $Y_i = \beta_o^{*} + \beta_1^{*} X_1 + \ldots + \beta_q^{*} X_q + \epsilon^{*}$

แล้วประมาณค่า β_i^* , i = 0, 1, ... ,q เหล่านี้ โดยวิธีกำลังสองน้อยที่สุด สมมุติว่าได้

$$\hat{\beta}^* = \begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \\ \vdots \\ \hat{\beta}_q^* \end{bmatrix} .$$

นำค่า $\hat{\beta}^*$ ที่ได้มาคำนวณหาค่า SSR_1

ซึ่ง $SSR_1 = \hat{\beta}^{*\prime} X_1' Y$ โดยที่ $X = [X_1 | X_2]$

X_1 เป็นเมตริกซ์ที่ประกอบด้วย $q+1$ หลักแรกของ X

แบบที่ 4 $H_o : \beta_k = 0 , \quad 1 \leq k \leq p$

ในการทดสอบเราจะถือเสมือนว่า β_k ถูกดึงเข้ามาอยู่ในตัวแบบเป็นตัวสุดท้าย

สร้างตารางวิเคราะห์ความแปรปรวน

ตารางที่ 3.5

ANOVA

Source of variation	d.f.	Sum of Squares	Mean Squares	F
All p variables	p	$\hat{\beta}' X' Y = SSR$		
Due to β_k	1	$\hat{\beta}^{*\prime} X_1' Y = SSR_1$		
Difference	p-1	$SSR - SSR_1$	$MSR_1 = \frac{SSR - SSR_1}{p-1}$	$\frac{MSR_1}{MSE}$
Residual	n-p-1	$SSE = YY - SSR$	$MSE = SSE/_{n-p-1} = s^2$	
Total	n-1	YY		

ในการคำนวณหาค่า SSR_1 และการสรุปผลการทดสอบ F ใช้วิธีการเดียวกับแบบที่ 3 (ใช้ t - test ก็ได้)

3.4 <u>สหสัมพันธ์</u> (Correlation)

3.4.1 <u>สัมประสิทธิ์สหสัมพันธ์เชิงเส้นอย่างง่าย</u> (Simple linear Correlation Coefficient)

สหสัมพันธ์เป็นการศึกษาถึงความสัมพันธ์ของตัวแปร 2 ตัว นั่นคือ สหสัมพันธ์จะแสดงอัตราที่ตัวแปร 2 ตัว ผันแปรอย่างเกี่ยวข้องกัน ค่าที่แสดงลักษณะนี้จะเรียกว่า สัมประสิทธิ์สหสัมพันธ์เชิงเส้นอย่างง่าย

ถ้าให้ r_{12} เป็นสัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้จากค่าสังเกตของ X_1 และ X_2 อย่างละ n ค่า แล้ว

$$r_{12} = \sum_{i=1}^{n} x_1 x_2 \Big/ \sqrt{(\sum_{i=1}^{n} x_1^2)(\sum_{i=1}^{n} x_2^2)}$$

โดยที่ $x_1 = X_1 - \bar{X}_1$

$x_2 = X_2 - \bar{X}_2$

ค่าของ r_{12} จะอยู่ระหว่าง -1 และ $+1$ ค่า r ที่เป็น + แสดงว่า X_1 และ X_2 มีความสัมพันธ์ไปในทางเดียวกัน คือ ถ้า X_1 เพิ่มแล้ว X_2 จะเพิ่มด้วย แต่ถ้าค่า r เป็น — ความสัมพันธ์ของตัวแปรทั้งสองจะเป็นไปในลักษณะตรงกันข้าม คือ ถ้า X_1 เพิ่ม X_2 จะลด และในทำนองเดียวกัน ถ้า X_1 ลด X_2 จะเพิ่ม ถ้า $r_{12} = 0$ แสดงว่า X_1 และ X_2 ไม่มีความสัมพันธ์กัน

3.4.2 สัมประสิทธิ์สหสัมพันธ์เพียงบางส่วน (Partial Correlation Coefficient)

สมมุติว่ามีตัวแปรอยู่หลายตัว เราต้องการจะหาความสัมพันธ์ของตัวแปรคู่หนึ่ง โดยที่ตัวแปรตัวอื่น ๆ คงที่ เช่น ในกรณีที่มีตัวแปร 3 ตัว เราสามารถหาสหสัมพันธ์อย่างง่ายได้ 3 ตัว คือ r_{12}, r_{13}, r_{23} แต่ถ้าเราต้องการจะหาความสัมพันธ์ระหว่างตัวแปรตัวที่ 1 กับตัวแปรตัวที่ 2 โดยที่ตัวแปรตัวที่ 3 คงที่ เราจะหาได้จากสูตร

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$r_{12.3}$ จะแสดงความสัมพันธ์ระหว่างตัวแปรตัวที่ 1 กับตัวที่ 2 ในขณะที่ตัวที่ 3 คงที่

ถ้ามี ตัวแปร 4 ตัว แล้วเราจะหาความสัมพันธ์ระหว่างตัวแปรตัวที่ 1 กับตัวที่ 2 โดยที่ตัวแปรตัวที่ 3 และ ตัวแปรตัวที่ 4 คงที่ จะหาได้จากสูตร

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}}$$

หรือ

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}}$$

การหาความสัมพันธ์ระหว่างตัวแปรคู่หนึ่ง เมื่อตัวแปรตัวอื่น ๆ คงที่ดังได้กล่าวแล้วอย่างนี้ เรียกว่า การหาความสัมพันธ์เพียงบางส่วน (Partial correlation)

3.5 การเลือกสมการถดถอยที่ดีที่สุด

สมมติว่าเราต้องการที่จะหาสมการถดถอยเชิงเส้น สำหรับตัวแปรตามซึ่งขึ้นอยู่กับตัวแปรอิสระ หรือตัวทำนาย $X_1, X_2, \ldots, X_k$ นั้น ในการเลือกสมการเรามักจะพิจารณาถึงจำนวนตัวแปรอิสระที่จะใช้ในการทำนาย และค่าใช้จ่ายที่จะใช้ในการเก็บตัวอย่างและประมวลผล ถ้าจำนวนตัวแปรอิสระที่ใช้ในการทำนายมากก็จะทำให้สามารถใช้ทำนายค่าของ Y ได้ใกล้เคียงยิ่งขึ้น แต่ถ้าจำนวนตัวแปร X มากเท่าไรก็จะเป็นการเพิ่มค่าใช้จ่ายมากเท่านั้น นอกจากนั้นยังทำให้การคำนวณยากยิ่งขึ้นด้วย ดังนั้น ถ้าพิจารณาในแง่เหล่านี้แล้ว เราก็ต้องการให้มีตัวแปรน้อยที่สุดคือ ให้มีเท่าที่จำเป็นเท่านั้น

ในการเลือกสมการถดถอยให้ได้สมการที่ดีที่สุดนั้น มีวิธีการทางสถิติหลายวิธีที่จะใช้ ทั้งนี้ในการพิจารณาว่าสมการไหนจะดีที่สุดย่อมขึ้นอยู่กับการตัดสินใจของบุคคล (personal judgement) ที่ดำเนินงานสถิตินั้น ๆ ด้วย

วิธีการทางสถิติที่จะพิจารณาถึงในที่นี้ จะพิจารณาถึง 4 วิธี คือ

(1) การถดถอยที่จะเป็นไปได้ทุกกรณี (All Possible Regressions)

(2) วิธีการกำจัดออกทีละตัว (The Backward Elimination Procedure)

(3) วิธีการเลือกเข้ามาทีละตัว (The Forward Selection Procedure)

(4) วิธีการถดถอยเป็นขั้นตอน (Stepwise Regression Procedure)

(1) <u>การถดถอยที่จะเป็นไปได้ทุกกรณี</u>

วิธีนี้เป็นการพิจารณาสมการถดถอยทุกสมการที่ประกอบด้วยตัวแปร X ต่าง ๆ แล้วนำมาเปรียบเทียบกันดูว่า สมการไหนจะดีที่สุด และเหมาะสมที่สุด กล่าวคือ

ถ้า $Y = f(X_1, X_2, \ldots, X_k)$ ซึ่งแสดงว่ามี X อยู่ k ตัว ที่มีอิทธิพลต่อ Y หาสมการถดถอยที่จะเป็นไปได้ทุกกรณี แล้วแบ่งสมการถดถอยออกเป็นกลุ่ม ๆ ตามจำนวนตัวแปร นั่นคือ สมการถดถอยที่อยู่ในกลุ่มเดียวกันจะประกอบด้วยจำนวนตัวแปรเท่ากัน

กลุ่มที่ 1 : $Y = \beta_o + \epsilon$ มีอยู่ 1 สมการ

กลุ่มที่ 2 : $Y = \beta_o + \beta_i X_i + \epsilon, \quad i = 1, 2, \ldots, k$

กลุ่มที่ 2 จะมีสมการถดถอยอยู่ $\binom{k}{1}$ สมการ

กลุ่มที่ 3 : $Y = \beta_o + \beta_i X_i + \beta_j X_j + \epsilon, \quad i \neq j$

$i = 1, 2, \ldots, k$

$j = 1, 2, \ldots, k$

กลุ่มที่ 3 จะมีสมการถดถอยอยู่ $\binom{k}{2}$ สมการ

กลุ่มที่ 4 .....

กลุ่มสุดท้าย : $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$

ซึ่งจะมีสมการถดถอยอยู่ $\binom{k}{k} = 1$ สมการ

ดังนั้น สมการถดถอยทั้งหมด $= \binom{k}{0} + \binom{k}{1} + \ldots + \binom{k}{k} = 2^k$

นั่นคือ จะมีสมการที่เราจะต้องพิจารณาอยู่ 2^k สมการ เช่น ถ้า $k = 10$ เราจะได้สมการถดถอยถึง $2^{10} = 1024$ สมการ

จากสมการที่ได้ในแต่ละกลุ่มเราจะเปรียบเทียบค่า R^2 ถ้าสมการไหนให้ค่า R^2 สูงสุด เราจะเลือกสมการนั้นของแต่ละกลุ่มออกมา เมื่อได้สมการที่ให้ค่า R^2 สูงสุดของแต่ละกลุ่มแล้ว จะนำสมการดังกล่าวมาเปรียบเทียบค่า R^2 กันอีกครั้งหนึ่งเพื่อจะศึกษาดูว่า จะเลือกสมการใดจึงจะดีที่สุด

<u>ตัวอย่าง</u> ถ้าตัวแปรตาม Y ขึ้นอยู่กับตัวแปรอิสระ X 4 ตัว คือ X_1, X_2, X_3, X_4 หาสมการถดถอยที่จะเป็นไปได้ทุกกรณี แล้วแบ่งออกเป็นกลุ่ม ๆ ได้ดังนี้

กลุ่มที่ 1 : $Y = \beta_0 + \epsilon$

กลุ่มที่ 2 เป็นสมการที่ประกอบด้วยตัวแปร 1 ตัว มี 4 สมการ

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

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

$$Y = \beta_0 + \beta_3 X_3 + \epsilon$$

$$Y = \beta_0 + \beta_4 X_4 + \epsilon$$

กลุ่มที่ 3 เป็นสมการที่ประกอบด้วยตัวแปร 2 ตัว มี 6 สมการ คือ

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

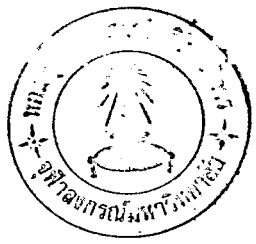
$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \epsilon$$

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

$$Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \epsilon$$

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$



กลุ่มที่ 4 เป็นสมการที่ประกอบด้วยตัวแปร 3 ตัว มี 4 สมการ คือ

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

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \epsilon$$

$$Y = \beta_o + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

กลุ่มที่ 5 เป็นสมการที่ประกอบด้วยตัวแปร 4 ตัว มี 1 สมการ คือ

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

ในแต่ละกลุ่มจะเลือกสมการที่ให้ค่า R^2 สูงสุด สมมุติว่าได้ดังนี้

กลุ่มที่	ตัวแปรในสมการ	ค่า R^2
2	$\hat{Y} = f(X_4)$	67.5 %
3	$\hat{Y} = f(X_1, X_2)$	97.9 %
	$\hat{Y} = f(X_1, X_4)$	97.2 %
4	$\hat{Y} = f(X_1, X_2, X_4)$	98.234 %
5	$\hat{Y} = f(X_1, X_2, X_3, X_4)$	98.237 %

จะเห็นว่า หลังจากดึงตัวแปรมา 2 ตัวแล้ว ค่า R^2 จะเพิ่มขึ้นเพียงเล็กน้อย สมการที่ควรพิจารณาจึงเป็นสมการในกลุ่มที่ 3 แต่จะเลือกสมการไหนจึงจะดีที่สุด ถ้าพิจารณาจาก R^2 เห็นว่าควรเลือก $\hat{Y} = f(X_1, X_2)$ แต่ถ้าพิจารณาจากกลุ่มที่ 2 แล้วจะเห็นว่า ถ้าพิจารณาตัวแปรเพียงตัวเดียวแล้ว ตัวแปร X_4 จะมีความสำคัญสูงสุดกับ Y ถ้าพิจารณาในแง่นี้ บางคนอาจจะเลือกสมการ $\hat{Y} = f(X_1, X_4)$ ก็ได้

เมื่อพิจารณาวิธีการในการเลือกสมการถดถอยแบบวิธีการพิจารณาสมการถดถอยที่จะเป็นไปได้ทุกกรณี จะเห็นว่า จะต้องพิจารณาสมการถดถอยเป็นจำนวนมากยิ่งถ้าตัวแปร X มากแล้ว ก็จะต้องศึกษาสมการถดถอยเพิ่มขึ้นอีกมาก ซึ่งบางครั้งก็ไม่จำเป็นนักที่จะต้องพิจารณาทุก ๆ สมการ การคำนวณโดยวิธีนี้ ถ้าตัวแปร X มาก การใช้เครื่องจักรคำนวณก็ยิ่งจำเป็นมาก และทำให้เสียเวลาเครื่องคำนวณมากด้วย

(2) <u>วิธีการกำจัดออกทีละตัว</u>

วิธีนี้เป็นวิธีที่ปรับปรุงมาจากการถดถอยที่จะเป็นไปได้ทุกกรณี โดยที่วิธีนี้จะไม่พิจารณาสมการถดถอยที่จะเป็นไปได้ทุกกรณี แต่จะพิจารณาเฉพาะสมการถดถอยที่ประกอบด้วยตัวแปร X ทั้งหมด เพียงสมการเดียว แล้วใช้วิธีกำจัดตัวแปร X ที่มีความสำคัญต่อ Y น้อยที่สุดออกจากสมการถดถอยทีละตัว วิธีการของการกำจัดออกทีละตัวมีดังนี้

ขั้นที่ 1 หาสมการถดถอยที่ประกอบด้วยตัวแปร X_i ทุกตัว

ขั้นที่ 2 คำนวณหา การทดสอบ F เพียงบางส่วน (Partial F - test) ของ b_i, $i = 1, 2, \ldots, k$ โดยถือว่า X_i แต่ละตัวถูกรวมเข้าไปในตัวแบบ (model) เป็นเทอมสุดท้าย

ขั้นที่ 3 กำหนดค่า F_o ซึ่งเป็นค่าที่บอกระดับนัยสำคัญ

ถ้า F_L เป็นค่า F เพียงบางส่วน (Partial F) ของ b_i ซึ่งมีค่าน้อยที่สุด และ $F_L < F_o$ แล้ว ให้ตัด X_L ซึ่งตรงกัน (corresponse) กับ F_L ออกจากตัวแบบ แล้วหาสมการถดถอยของตัวแปรที่เหลือต่อจากนั้นจึงเริ่มไปทำขั้นที่ 2 ใหม่

แต่ถ้า $F_L \geq F_o$ ก็ให้ถือว่า ตัวแบบที่เราเริ่มต้นนั้น ถูกต้องแล้ว จะได้ตัวแบบนั้นเป็นตัวแบบที่ต้องการ

วิธีการกำจัดออกทีละตัวนี้ นับว่าเป็นวิธีการที่ดีวิธีหนึ่งสำหรับนักสถิติที่ต้องการจะทดสอบสมการโดยเริ่มตั้งแต่สมการนั้นประกอบด้วยตัวแปรทุกตัว และเมื่อเปรียบเทียบวิธีนี้กับวิธีการถดถอยที่จะเป็นไปได้ทุกกรณีแล้ว จะเห็นว่า วิธีนี้ประหยัดเวลา และแรงงานที่ใช้ในการคำนวณมากกว่า และถ้าใช้เครื่องจักรคำนวณช่วยในการประมวลผล ก็จะประหยัดเวลาเครื่องจักรคำนวณมากกว่าวิธีแรกมาก

<u>ตัวอย่างสำหรับวิธีการกำจัดออกทีละตัว</u>

สมมุติมีตัวแปรอยู่ 4 ตัว คือ Y, X_1, X_2, X_3 โดยที่ Y เป็นตัวแปรตาม และ X_1, X_2, X_3 เป็นตัวแปรอิสระ มีค่าสังเกตอยู่ 15 จำนวน

<u>ขั้นที่ 1</u> หาสมการถดถอยที่ประกอบด้วยตัวแปร X_i ทุกตัว ได้เป็น

$$\hat{Y} = -7.84 + .220X_1 + .219X_2 + .424X_3$$

<u>ตารางที่ 3.6</u>

<u>ANOVA</u>

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	3	1292.4	430.8	10.46
Residual	11	453.2	$41.2 = s^2$	
Total (Corrected)	14	1745.6		

ขั้นที่ 2 คำนวณหาค่าการทดสอบ F เพียงบางส่วน

$$\frac{SS(b_1/b_o, b_2, b_3)}{s^2} = 1.87$$

$$\frac{SS(b_2/b_o, b_1, b_3)}{s^2} = 3.89$$

$$\frac{SS(b_3/b_o, b_1, b_2)}{s^2} = 11.68$$

ได้ $F_L = 1.87$

ขั้นที่ 3 กำหนด F_o ที่ระดับนัยสำคัญ 5% ได้ $F_o(1,11,0.95) = 4.84$

จะเห็นว่า $F_L < F_o$ ดังนั้นจึงต้องตัด X_1 ออกจากตัวแบบ เมื่อตัด X_1 ออกแล้ว สมการถดถอยจะเป็น

$$\hat{Y} = 6.08 + .276X_2 + .425X_3$$

ตารางที่ 3.7

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F.
Regression	2	1221.4	610.7	13.97
Residual	12	524.2	$43.7 = s^2$	
Total (corrected)	14	1745.6		

ขั้นที่ 4 คำนวณค่าทดสอบ F เพียงบางส่วน

$$\frac{SS(b_2 / b_o, b_3)}{s^2} = 6.926$$

$$\frac{SS(b_3 / b_o, b_2)}{s^2} = 10.952$$

ได้ $F_L = 6.926$

จากตาราง $F_o(1, 12, 0.95) = 4.75$

นั่นคือ $F_L \geq F_o$ แสดงว่าไม่สามารถตัดตัวแปรใดออกจากตัวแบบได้อีก

ดังนั้น สมการถดถอยที่ต้องการ คือ

$$\hat{Y} = 6.08 + .276X_2 + .425X_3$$

(3) วิธีการเลือกเข้ามาทีละตัว

วิธีนี้เป็นวิธีที่มีขั้นตอนในการหาเพื่อให้ได้สมการถดถอยที่ต้องการตรงกันข้ามกับวิธีกำจัดออกทีละตัว กล่าวคือ วิธีการเลือกเข้ามาทีละตัวนั้น จะพยายามหาตัวแปร X ที่มีความสำคัญ หรือมีความสัมพันธ์กับ Y มากที่สุดใส่เข้าไปในสมการทีละตัว จนได้สมการที่ต้องการ ในการหาตัวแปรที่เหมาะสมใส่เข้าไปในสมการจะใช้สัมประสิทธิ์สหสัมพันธ์เพียงบางส่วน (Partial correlation coefficient) เป็นตัววัดความสำคัญของตัวแปรที่ยังมิได้อยู่ในสมการ

วิธีการเลือกเข้ามาทีละตัวมีดังนี้

ถ้า Y เป็นตัวแปรที่ขึ้นอยู่กับตัวแปร X k ตัวคือ $X_1, X_2, \ldots, X_k$

ขั้นที่ 1 หาสหสัมพันธ์ (correlation) ระหว่าง X_i กับ Y ถ้าตัวแปร X_1 มีสหสัมพันธ์กับ Y มากที่สุด ให้สร้างสมการถดถอย

$$\hat{Y} = f(X_1)$$

ขั้นที่ 2 หาสหสัมพันธ์เพียงบางส่วน (Partial correlation) ระหว่าง X_j ที่เหลือ กับ Y (เมื่อมี X_1 อยู่ในสมการแล้ว) นั่นคือเป็นการหาสหสัมพันธ์ระหว่าง ส่วนที่เหลือ (residuals) จากสมการถดถอย $\hat{Y} = f(X_1)$ กับส่วนที่เหลือจากสมการถดถอย $\hat{X}_j = f_j(X_1)$ X_j ที่ให้ค่าสหสัมพันธ์เพียงบางส่วน กับ Y สูงที่สุด จะถูกเลือกเข้ามาอยู่ในสมการเป็นตัวต่อมา (สมมุติว่าเป็น X_2) จะได้สมการถดถอยใหม่เป็น

$$\hat{Y} = f(X_1, X_2)$$

ขั้นที่ 3 พิจารณาค่าทดสอบ F เพียงบางส่วน (Partial F - test) สำหรับตัวแปรที่เพิ่งถูกดึงเข้ามาอยู่ในสมการ ถ้ามีนัยสำคัญให้ถือว่าตัวแปรตัวที่เพิ่งถูกดึงเข้ามาอยู่ในสมการนั้น มีความสำคัญต่อ Y ให้คงไว้ แล้วเลือกตัวแปรตัวใหม่ต่อไปตามวิธีการในขั้นที่ 2 แต่ถ้าไม่มีนัยสำคัญ ให้ตัดตัวแปรตัวสุดท้ายที่เพิ่งถูกดึงเข้ามาอยู่ในสมการออก วิธีการเลือกเข้ามาทีละตัวจะหยุดแค่นี้ ให้ถือว่าสมการที่เพิ่งตัดตัวแปรที่ไม่มีความสำคัญต่อ Y ออกนี้เป็นสมการที่ต้องการ

เมื่อเปรียบเทียบกับวิธีการกำจัดออกทีละตัวแล้ว จะเห็นว่าวิธีการเลือกเข้ามาทีละตัวเป็นการหลีกเลี่ยงที่จะไปคำนวณเกี่ยวกับ X ที่ไม่มีความสำคัญสำหรับ Y ทำให้ประหยัดเวลาในการคำนวณมากกว่า 2 วิธีแรก แต่วิธีการเลือกเข้ามาทีละตัวมิได้ตรวจสอบดูว่าในการดึงตัวแปรอิสระเข้ามาใหม่นั้น มันอาจจะมีผลหรือมีอิทธิพลต่อตัวแปรอื่นที่ถูกดึงเข้ามาอยู่ในสมการถดถอยก่อนแล้วก็ได้ วิธีการที่เรียกว่า

วิธีการถดถอยเป็นขั้นตอน (Stepwise regression procedure) ได้แก้ไขข้อบกพร่องอันนี้

ตัวอย่างสำหรับวิธีการเลือกเข้ามาทีละตัว

ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลชุดเดียวกับที่ใช้ในวิธีการกำจัดออกทีละตัว นั่นคือ Y เป็นตัวแปรที่ขึ้นอยู่กับตัวแปร X_1, X_2, X_3

ขั้นที่ 1 . หาสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_i, $i = 1,2,3$ กับ Y

ให้ r_{i4} แทนสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_i กับ Y

ได้ $r_{14} = .4574$, $r_{24} = .6523$, $r_{34} = .7255$

ตัวแปรที่ถูกเลือกตัวแรก คือ X_3 สร้างสมการถดถอยได้

$$\hat{Y} = 13.8085 + .5479 X_3$$

ตารางที่ 3.8

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	1	918.7	918.7	14.44
Residual	13	826.9	63.6	
Total (corrected)	14	1745.6		

ขั้นที่ 2 หาค่าสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนของตัวแปรที่ไม่อยู่ในสมการถดถอยได้ $r_{14.3} = .5053$, $r_{24.3} = .5992$ ดังนั้นตัวแปรที่ถูกเลือกเป็นตัวต่อมาคือ X_2 ได้สมการถดถอยใหม่เป็น

$$\hat{Y} = 6.08 + .425 X_3 + .276 X_2$$

ตารางที่ 3.9

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	2	1221.4	610.7	13.97
Residual	12	524.2	$43.7 = s^2$	
Total (corrected)	14	1745.6		

เนื่องจาก $\frac{SS(b_2 / b_o, b_3)}{s^2} = 6.926$ มากกว่า

$F(1, 12, 0.95) = 4.75$ แสดงว่า X_2 มีความสำคัญต่อ Y ให้คงไว้ในสมการ

ขั้นที่ 3 หาสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนของตัวแปรที่เหลือได้ $r_{14.23} = .3678$ ได้สมการถดถอยใหม่เป็น

$$\hat{Y} = -7.84 + .220 X_1 + .219 X_2 + .424 X_3$$

ตารางที่ 3.10

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	3	1292.4	430.8	10.46
Residual	11	453.2	41.2 = s^2	
Total (corrected)	14	1745.6		

$\frac{SS(b_1 / b_o, b_2, b_3)}{s^2} = 1.87$ น้อยกว่า $F(1,11,0.95) = 4.84$

แสดงว่า X_1 ไม่มีความสำคัญต่อ Y พอ จึงตัด X_1 ออกจากตัวแบบ ดังนั้นสมการถดถอยที่ต้องการ คือ

$$\hat{Y} = 6.08 + .425 X_3 + .276 X_2$$

(4) วิธีการถดถอยเป็นขั้นตอน

วิธีการถดถอยเป็นขั้นตอนเป็นวิธีที่ปรับปรุงมาจากวิธีการเลือกเข้ามาทีละตัว กล่าวคือ ในการดึงตัวแปรใหม่เข้ามาอยู่ในสมการถดถอยแต่ละครั้งจะมีการตรวจสอบดูว่าตัวแปรที่ดึงเข้ามาใหม่นี้มีอิทธิพลต่อตัวแปรที่อยู่ในสมการแล้วหรือไม่ เพราะว่าตัวแปรบางตัวที่เราเลือกเข้ามาอยู่ในตัวแบบในตอนแรก อาจจะไม่จำเป็น ถ้าเราดูความสัมพันธ์ของมันกับตัวแปรที่เราดึงเข้ามาใหม่ ในการตรวจสอบดังกล่าวนี้จะใช้หลักของการทดสอบค่า F เพียงบางส่วน (Partial F - test) โดยถือว่าตัวแปรแต่ละตัวนั้นถูกใส่เข้าไปในตัวแบบเป็นตัวสุดท้าย ตัวใดที่ไม่มีนัยสำคัญ (non - significant) ให้ตัดออกจากตัวแบบ คงไว้เฉพาะตัวที่มีนัยสำคัญ (significant) วิธีการถดถอยเป็นขั้นตอนจะหยุดก็ต่อเมื่อเราไม่สามารถเพิ่มหรือลดตัวแปรใด ๆ ในตัวแบบได้อีก

วิธีการถดถอยเป็นขั้นตอนแสดงเป็นลำดับขั้นได้ดังนี้

ขั้นที่ 1 หาสหสัมพันธ์ระหว่าง X_i, i = 1, 2,...., k กับ Y X_i ตัวใดที่มีสหสัมพันธ์กับ Y สูงสุดจะถูกเลือกเข้ามาอยู่ในตัวแบบเป็นอันดับแรก

ขั้นที่ 2 หาการทดสอบค่า F เพียงบางส่วน (Partial F - test) ของตัวแปรทุก ๆ ตัว โดยถือว่าตัวแปรนั้น ๆ ถูกใส่เข้าไปในตัวแบบเป็นตัวสุดท้าย ตัวใดที่ไม่มีนัยสำคัญให้ตัดออกจากตัวแบบ

ขั้นที่ 3 หา สหสัมพันธ์เพียงบางส่วน (partial correlation) ของตัวแปรที่ยังไม่ถูกใส่เข้าไปในตัวแบบกับ Y ตัวที่ให้ค่าสหสัมพันธ์เพียงบางส่วนสูงที่สุด จะถูกดึงเข้ามาอยู่ในตัวแบบแล้วย้อนกลับไปทำ ขั้นที่ 2 ใหม่

ขั้นที่ 4 วิธีการถดถอยเป็นขั้นตอนจะหยุดเมื่อเราไม่สามารถเพิ่มหรือลดตัวแปรใด ๆ เข้าไปในตัวแบบได้อีก ก็จะได้สมการที่ดีที่สุดสำหรับวิธีการนี้

ในการวิจัยครั้งนี้ ได้เลือกใช้วิธีการถดถอยเป็นขั้นตอนในการวิเคราะห์ข้อมูล เพื่อให้ได้สมการถดถอยตามต้องการ เหตุที่เลือกใช้วิธีนี้ก็เพราะว่าได้พิจารณาแล้วเห็นว่า วิธีการถดถอยเป็นขั้นตอน น่าจะใช้ได้ดีกว่าวิธีการอื่น ๆ ที่กล่าวถึง

ตัวอย่างสำหรับวิธีการถดถอยเป็นขั้นตอน

ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลชุดเดียวกับที่ใช้ในวิธีการกำจัดออกทีละตัว นั่นคือ Y เป็นตัวแปรตามที่ขึ้นอยู่กับ X_1, X_2, X_3 ซึ่งเป็นตัวแปรอิสระ และต่างก็มีจำนวนค่าสังเกตอยู่ 15 จำนวน

กำหนดให้ระดับ F สำหรับดึงตัวแปรเข้า และตัดตัวแปรออกจากตัวแบบเป็น 3.28

ขั้นที่ 1 หา สัมประสิทธิ์สหสัมพันธ์ระหว่าง X_i, $i = 1,2,3$ กับ Y

ให้ r_{i4} แทนสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_i กับ Y

ได้ $r_{14} = .4574$, $r_{24} = .6523$, $r_{34} = .7255$

ดังนั้น ตัวแปรที่ถูกเลือกเป็นตัวแรก คือ X_3

• สร้างสมการถดถอยได้

$$\hat{Y} = 13.8085 + .5479 X_3$$

ตารางที่ 3.11

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	1	918.7	918.7	14.44
Residual	13	826.9	63.6	
Total (corrected)	14	1745.6		

ค่าสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนของตัวแปรที่ไม่อยู่ในสมการถดถอยเป็นดังนี้

$$r_{14.3} = .5053 \quad , \quad r_{24.3} = .5992$$

ขั้นที่ 2 ดึงตัวแปร X_2 เข้ามาเป็นตัวที่สอง ได้สมการถดถอยเป็น

$$\hat{Y} = 6.08 + .425 X_3 + .276 X_2$$

ตารางที่ 3.12

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	2	1221.4	610.7	13.97
Residual	12	524.2	$43.7 = s^2$	
Total (corrected)	14	1745.6		

ค่าทดสอบ F เพียงบางส่วน

$$\frac{SS(b_2 \,/\, b_0, b_3)}{s^2} = 6.926 \geq 3.28$$

$$\frac{SS(b_3 \,/\, b_0, b_2)}{s^2} = 10.925 \geq 3.28$$

ค่าสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วน $r_{14.23} = .3678$

ขั้นที่ 3 ดึงตัวแปร X_1 เข้ามาเป็นตัวที่สาม ได้สมการถดถอยเป็น

$$\hat{Y} = -7.84 + .220\,X_1 + .219\,X_2 + .424\,X_3$$

ตารางที่ 3.13

ANOVA

Source of variation	d.f.	Sum of squares	Mean squares	F
Regression	3	1292.4	430.8	10.46
Residual	11	453.2	$41.2 = s^2$	
Total (corrected)	14	1745.6		

$$\frac{SS(b_1 / b_o, b_2, b_3)}{s^2} = 1.87 \not> 3.28$$

$$\frac{SS(b_2 / b_o, b_1, b_3)}{s^2} = 3.89 > 3.28$$

$$\frac{SS(b_3 / b_o, b_1, b_2)}{s^2} = 11.68 > 3.28$$

<u>ขั้นที่ 4</u> ตัดตัวแปร X_1 ออกจากตัวแบบ

ดังนั้นได้สมการถดถอยที่ต้องการ คือ

$$\hat{Y} = 6.08 + .425 X_3 + .276 X_2$$