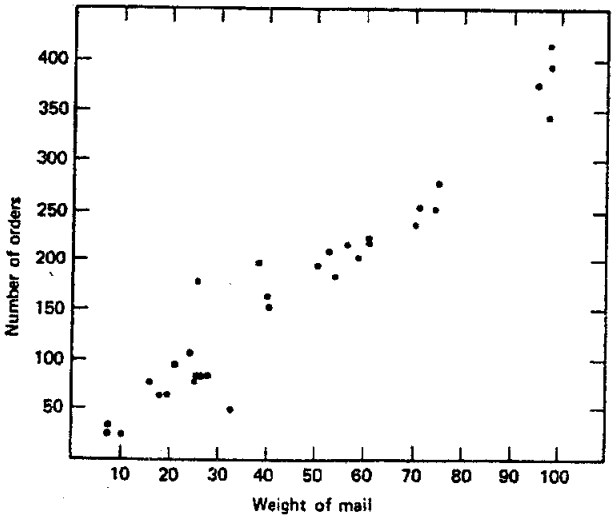
บทที่ 7
วิธีการวิเคราะห์การถดถอย

เทคนิควิธี regression analysis นี้พยายามที่จะทำนายอนาคต โดยวิธีค้นพบและวัดค่าปัจจัยอิสระที่สำคัญ ๆ หลาย ๆ ตัว และผลกระทบที่มีต่อค่าตัวแปรในการพยากรณ์และเพื่อความสัมพันธ์ในสิ่งที่สนใจ แล้วใช้ผลลัพธ์ที่ได้ไปเพื่อการพยากรณ์ เพราะว่าค่าใช้จ่ายที่สูงกว่าวิธีนี้เหมาะสำหรับการวางแผนระยะยาวและในสถานการณ์ซึ่งต้องการความแม่นยำของค่าพยากรณ์ที่สูงกว่า เทคนิคนี้แบ่งเป็น simple regression ซึ่งเป็นเทคนิคที่ไม่เพียงสมมติว่ามีรูปแบบพื้นฐานเกิดขึ้นเท่านั้น แต่ตัวแบบของรูปแบบพื้นฐานนั้นเป็นเชิงเส้น และอีกวิธีหนึ่งคือ multiple regression โดย multiple regression มีตัวแปรตามหนึ่งตัวที่จะทำนาย แต่มีตัวแปรอิสระ 2 ตัวหรือมากกว่า การใช้ simple regression เริ่มจากสมมติฐานที่ว่าความสัมพันธ์พื้นฐานระหว่างสองตัวแปรเกิดขึ้น โดยตัวแปรอิสระสามารถใช้ในการทำนายค่าของตัวแปรตามบางตัว และสามารถแทนด้วยรูปแบบของฟังก์ชันเชิงคณิตศาสตร์ คือ $y = f(x)$ ค่าของ y เป็นค่าของตัวแปรตาม ซึ่งขึ้นอยู่กับค่าของตัวแปร x ซึ่งเป็นตัวแปรอิสระ ใน simple regression เป็นความสัมพันธ์เชิงเส้น ซึ่งเขียนได้เป็น $Y = a + bX$ จากรูปแบบทั่วไปนี้ ถ้า $X = 0$ จะได้ $Y = a$ ซึ่งค่าของ a ได้จากกราฟเส้นตรงที่ตัดแกน y (y-intercept) ใน multiple regression มีสมการคือ $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$ เมื่อ Y เป็นตัวแปรตาม และ $X_1, X_2, \ldots, X_k$ เป็นตัวแปรอิสระ

7.1 Simple Regression

ค่าพยากรณ์จะแสดงในรูปของฟังก์ชันของจำนวนปัจจัยที่แน่นอน เพื่อใช้ในการตัดสินใจผลลัพธ์ของการพยากรณ์แต่ละการพยากรณ์ บางการพยากรณ์ไม่จำเป็นต้องขึ้นอยู่กับเวลา ซึ่งยอมให้ใช้เพื่อการทำนายค่าแนวนอนในเทอมที่ยาว ๆ ดีกว่าที่จะใช้อนุกรมเวลา เช่นเดียวกับการพัฒนาตัวแบบเชิงเหตุผลให้ความเข้าใจดีกว่าในสถานการณ์ ซึ่งการทดลองด้วยการรวมตัวที่แตกต่างของปัจจัยนำเข้าที่จะศึกษาถึงผลกระทบต่อค่าพยากรณ์ตัวแบบเหตุผล โดยรูปแบบพื้นฐานมีอิทธิพลในอนาคตต่อการตัดสินใจในวันนี้ ตั้งแต่ผลของการตัดสินใจกระทำในวันนี้ไม่เกิดขึ้นในบางเวลา การพยากรณ์เชิงเหตุผลจะมีความเหมาะสมสำหรับเวลาในแนวนอนของ 3 เดือนถึง 2 ปี



รูปที่ 7.1 แสดงการพยากรณ์จำนวนใบสั่งซื้อจากน้ำหนักของจดหมาย

จากรูป 7.1 จะได้ว่าเมื่อน้ำหนักของจดหมายเป็น 0 จำนวนใบสั่งซื้อจะมีค่าเท่ากับ a และถ้า a เป็น 0 แสดงว่าไม่มีใบสั่งซื้อได้รับเข้ามาเลย และค่า b ในสมการ เรียกว่า สัมประสิทธิ์ของการถดถอย ซึ่งเป็นตัวชี้ถึงการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนแปลงไป 1 หน่วย ดังนั้นถ้าเราเปรียบเทียบจำนวนใบสั่งซื้อจาก 40 และ 41 ปอนด์ของน้ำหนักจดหมาย เราสามารถคาดคะเนการเพิ่มขึ้นของ b เท่าของจำนวนใบสั่งซื้อ ในกรณี 41 ปอนด์ในเทอมอื่น ๆ จะ หมายถึง สัมประสิทธิ์ของการถดถอย b เป็นค่าความชันของกราฟเส้นตรง

7.1.1 รูปแบบต่าง ๆ ของความสัมพันธ์เชิงฟังก์ชัน

เป็นที่รู้กันว่า GNP เพิ่มขึ้นตามกาลเวลา ซึ่งสามารถเขียนได้ในรูป GNP = f (time)

รูปแบบของฟังก์ชันมีความสัมพันธ์กันระหว่าง 2 ตัวแปร (คือ GNP และ Time) บางครั้งไม่จำเป็นต้องรวมเวลาเข้าไปด้วย เราสามารถรวมตัวแปรอื่น ๆ ในความสัมพันธ์เชิงเหตุผลทางเศรษฐศาสตร์ เช่น ความต้องการของสินค้า (ยอดขาย) จะขึ้นอยู่กับราคาของสินค้านั้น เช่น

Demand of product X = f (price of product X)

คือเมื่อราคาของสินค้า X เปลี่ยน ความต้องการในสินค้า X จะเปลี่ยนแปลงด้วย สินค้าส่วนใหญ่เมื่อราคาของสินค้าเพิ่มขึ้นปริมาณยอดขายจะลดลง ขณะที่ราคาของสินค้าลดลง ปริมาณยอดขายจะเพิ่มขึ้น เช่นเดียวกันความสัมพันธ์เชิงฟังก์ชันไม่ได้มีขีดจำกัดเพียงสองตัวแปรเท่านั้น ยอดขายของบริษัทมีอิทธิพลจากหลาย ๆ ปัจจัยด้วยกัน คือ ระดับของ GNP , ราคาของสินค้า , งบประมาณในการโฆษณา , งบประมาณของการวิจัยและพัฒนา (R & D) , ราคาของชิ้นส่วนประกอบ ดังนั้นจึงอาจเขียนได้เป็น

Sales of XYZ = f (GNP , prices , advertising , R & D , prices of substitutes , etc.)

เมื่อยอดขายขึ้นอยู่กับตัวแปรมากกว่าหนึ่งตัว ถ้ารูปแบบของความสัมพันธ์เหล่านี้เปลี่ยนแปลงในหลาย ๆ ตัวแปร ยอดขายจะถูกค้นพบ มันสามารถที่จะใช้สำหรับการพยากรณ์ จุดประสงค์คือเราจะตรวจสอบกระบวนการสำหรับแยกแยะความสัมพันธ์และวัดค่าออกมา ดังนั้นรูปแบบที่เชื่อมต่อทางซ้ายและขวาของสมการสามารถแยกจาก ปัจจัยสุ่มได้ โดยตัวแปรทางซ้ายมือ แทนด้วย Y เรียกว่า ตัวแปรตาม ส่วนทางขวามือ เรียกว่า ตัวแปรอิสระ แทนด้วย X จุดประสงค์ของการพยากรณ์คือทำนายตัวแปรตามโดยค้นพบว่า ตัวแปรตามมีความสัมพันธ์หรือไม่กับตัวแปรอิสระตัวเดียวหรือมากกว่า เมื่อมีตัวแปรอิสระตัวเดียว เราใช้วิธีการของ simple regression จะมีความเหมาะสมกว่า ขั้นตอนในการพยากรณ์มี 2 ขั้นตอน คือ

ขั้นตอนที่ 1 พิจารณารูปแบบของความสัมพันธ์ระหว่าง 2 ตัวแปรว่าเป็นความสัมพันธ์เชิงเส้น , เอกซ์โปเนนเชียล , ควอดราติก (quadratic) หรือ cubic etc. ต้องทำการตัดสินใจออกมา

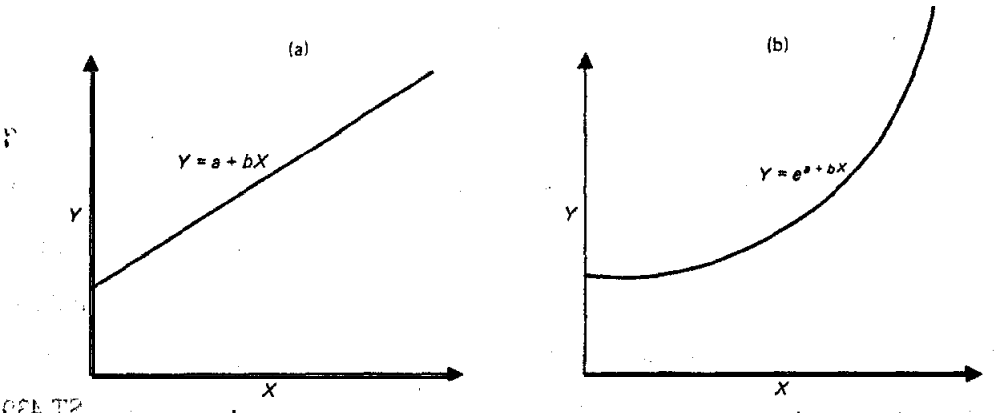
ขั้นตอนที่ 2 พารามิเตอร์ในตัวแบบต้องทำการประมาณค่า ในหลาย ๆ ตัวอย่างความสัมพันธ์ระหว่างสองตัวแปรซึ่งผู้จัดการเห็นว่าน่าจะเป็นความสัมพันธ์เชิงเส้น บางครั้งอาจจะไม่ปรากฏเป็นความสัมพันธ์เชิงเส้น เมื่อพล็อตกราฟโดยตรง แต่เมื่อแปลงค่าตัวแปรตัวใดตัวหนึ่งจะให้ผลเป็นตัวแปรใหม่ซึ่งจะมีความสัมพันธ์เชิงเส้นกับอีกตัวแปรหนึ่งที่ไม่ได้ทำการแปลงค่า กรณีง่าย ๆ นี้จะช่วยอธิบายตรงจุดนี้ได้ ตัวอย่างเช่น กรณี $W = AB^X$ เมื่อ A,B เป็นค่าคงที่ ตัวแปร W มีความสัมพันธ์ในรูปเอกซ์โปเนนเชียลมากกว่าความสัมพันธ์เชิงเส้น กับตัวแปร X เราสามารถแปลงให้อยู่ในรูปฟังก์ชันเชิงเส้น โดยการนำ logarithm มาแปลงดังนี้

$$W = AB^X$$

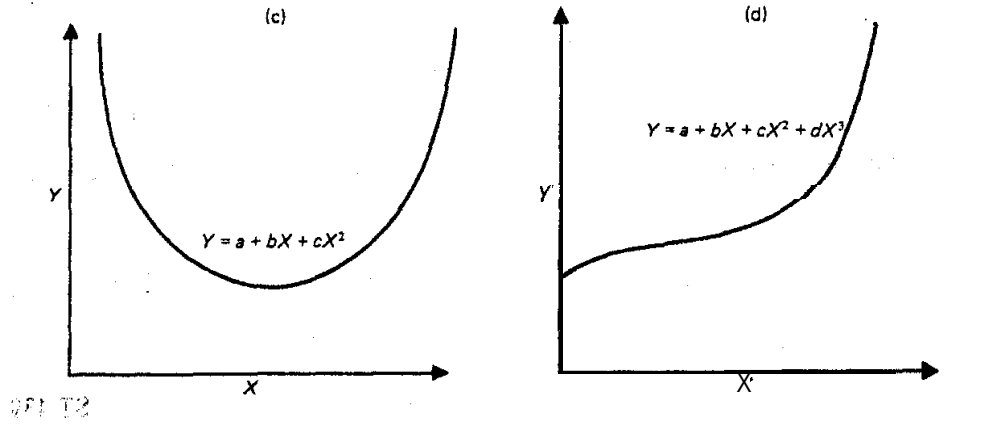
$$\log W = \log A + x \log B$$

เมื่อ $Y = \log W$, $a = \log A$, $b = \log B$ จะได้ $Y = a + bX$ ซึ่งเป็นรูปของฟังก์ชันเชิงเส้น การวิเคราะห์การถดถอยเชิงเส้น สามารถใช้ในการตัดสินใจเกี่ยวกับค่าของ a และ b ซึ่งสมการนี้สามารถใช้ในการพยากรณ์ค่า Y

อีกตัวอย่างหนึ่ง ถ้า W อยู่ในรูป $W = e^{a+bx}$ สามารถแปลงเป็นรูปฟังก์ชันเชิงเส้นได้ คือ $\ln W = a + bX$ เมื่อ $Y = \ln W$ จะได้ $Y = a + bX$ ซึ่งจะเป็นสมการความสัมพันธ์แบบเชิงเส้น



รูปที่ 7.2 แสดงกราฟของฟังก์ชันเชิงเส้นตรงและเอกซ์โปเนนเชียล



รูปที่ 7.3 แสดงกราฟของฟังก์ชันรูปแบบ Quadratic และ Cubic

7.1.2 การประมาณค่าพารามิเตอร์ a และ b

มีหลายวิธีการที่สามารถประมาณค่าของ a และ b จาก $Y = a + bX$ บางครั้งใช้วิธีง่ายๆ โดยพล็อตจุดของค่าสังเกตแล้วลากเส้นตรงให้ผ่านจุดเหล่านั้น แล้วค่า a และ b สามารถอ่านจากกราฟที่เขียนขึ้น ซึ่ง a เป็นค่าซึ่งเป็นจุดตัดแกน Y ส่วนค่า b สามารถหาค่าง่าย ๆ จากการเพิ่มค่า y ขึ้นแล้วค่า x เพิ่มขึ้นเท่าไร ให้นำผลต่างของค่า y ทั้งสองหารด้วยผลต่างของค่า x ที่เพิ่มขึ้น จึงเป็นค่า b ในกรณีที่มีข้อมูลจำนวนมาก ๆ การพล็อตกราฟจะเป็นการยากที่จะได้เส้นกราฟที่เหมาะสมดีที่สุด เราจึงควรใช้วิธีการประมาณค่า a และ b ด้วยวิธีทางคณิตศาสตร์ เรียกว่าวิธีกำลังสองน้อยที่สุด ดังนี้

จากสมการถดถอยอย่างง่าย คือ $Y_i = a + bX_i + \epsilon_i$

เมื่อ Y_i เป็นค่าของตัวแปรตามค่าที่ i

a , b เป็นพารามิเตอร์

X_i เป็นค่าของตัวแปรอิสระ ค่าที่ i

ϵ_i เป็นค่าความคลาดเคลื่อนจากการพยากรณ์

ภายใต้สมมติฐาน (assumption) ϵ_i เป็นตัวแปรเชิงสุ่มที่มี $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$, $E(\epsilon_i \epsilon_j) = 0$ เมื่อ $i \neq j$

ได้ $\hat{Y}_i = a + bX_i$

$\epsilon_i = e_i = Y_i - \hat{Y}_i$

ให้ $S = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2$

ต้องการให้ $S = \sum e_i^2$ มีค่าต่ำที่สุด โดย

$$\partial S / \partial a = -2 \sum (Y_i - a - bX_i) = 0$$

$$\partial S / \partial b = -2 \sum (Y_i - a - bX_i) X_i = 0$$

$$\sum Y_i - na - b \sum X_i = 0 \quad \text{.....(7.1)}$$

$$\sum X_i Y_i - a \sum X_i - b \sum X_i^2 = 0 \quad \text{.....(7.2)}$$

แก้สมการ (7.1) และ (7.2) ได้ค่า

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

$$a = \frac{\sum Y}{n} - b \frac{\sum X}{n} = \bar{Y} - b \bar{X}$$

ตัวอย่างที่ 7.1 X เป็นจำนวนสินค้า Y เป็นราคาของสินค้า แสดงดังตารางต่อไปนี้ จงหาสมการถดถอยเชิงเส้น

Y	X	X^2	Y^2	XY
8	3	9	64	24
11	2	4	121	22
16	5	25	256	80
15	7	49	225	105

$\sum Y = 50$ $\sum X = 17$ $\sum Y^2 = 666$ $\sum X^2 = 87$ $\sum XY = 231$ $\bar{Y} = 125$

$\bar{X} = 425$ ได้ $b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{4(231) - (50)(17)}{4(87) - (17)^2}$

$$b = \frac{74}{59} = 1.254$$

$$a = \bar{Y} - b\,\bar{X} = 12.5 - (1.254)(4.25) = 7.17$$

ได้สมการถดถอย คือ $\hat{Y}_i = 7.17 + 1.254\,X_i$

เมื่อ a = 7.17 , b = 1.254 ได้ค่า MSE ต่ำที่สุด โดย MSE = 7.195 เมื่อลองเปรียบเทียบกับค่า a , b ที่แตกต่างจากนี้ ค่า MSE จะโตกว่า โดย a = 10 , b = 1.2 จะมีค่า MSE = 11.22 และถ้า a = 10 , b = 1 MSE = 7.75

7.1.3 การวิเคราะห์ความแปรปรวนของการถดถอยเชิงเส้น

จากการพยากรณ์ความถูกต้องแม่นยำที่เกิดขึ้นสามารถดูได้จากการวิเคราะห์ความแปรปรวนของตัวแปร Y ได้ ซึ่งสามารถแบ่งความแปรปรวนออกเป็นสองส่วน คือ SSR = Sum Square Regression เป็นความแปรปรวนจากเส้นสมการถดถอยสามารถใช้อธิบายค่า Y ได้ และ SSE = Sum Square Error ซึ่งเป็นความแปรปรวนที่ไม่สามารถอธิบายได้

$$SST = SSR + SSE$$

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

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

7.1.3.1 การทดสอบสมมติฐาน ต้องการทดสอบว่าสัมประสิทธิ์ของ regression เป็น 0 หรือไม่ โดยตั้งสมมติฐาน คือ

$H_0 : b = 0 \quad H_1 : b \neq 0$

ก. กรณีทดสอบโดยใช้ F - test

ANOVA

SOV.	df	SS	MS	F
Regression	1	SSR	MSR = SSR / 1	MSR / MSE
Residual	n - 2	SSE	MSE = SSE / (n-2)	
Total	n - 1	SST		

ถ้า $F > F_{\alpha,(1,n-2)}$ เราจะปฏิเสธ H_0

ตัวอย่างที่ 7.2 จากตัวอย่างที่ 7.1 จงทดสอบสมมติฐาน $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$ โดยใช้ F - test

$$SST = \sum Y^2 - [(\sum Y)^2 / n] = 666 - [(50)^2/4] = 41$$

$SSR = b[\sum XY - \{(\sum X)(\sum Y)/n\}] = (1.254)[231 - \{(50)(17)/4\}]$

$= (1.254)(18.5) = 23.199$

$SSE = SST - SSR = 41 - 23.2 = 17.8$

ANOVA

SOV.	df	SS	MS	F
Regression	1	23.2	23.2	2.61
Residual	2	17.8	8.9	
Total	3	41.0		

$F_{0.05,(1,2)} = 18.51$ นั่นคือ $F_c < 18.51$ เรายอมรับ H_0

ข. กรณีทดสอบโดยใช้ t - test

<u>ทดสอบพารามิเตอร์ b</u> ซึ่งตั้งสมมติฐาน $H_0: b = b_0$ $H_1: b \neq b_0$

ตัวสถิติที่ใช้ทดสอบ คือ

$$t = \frac{b - b_0}{S/\sqrt{\sum(X_i - \bar{X})^2}}$$

เมื่อ ค่า S ได้จาก $\sqrt{MSE}$

ถ้า $|t_c| > t_{1-\alpha/2,n-2}$ เราจะปฏิเสธ H_0

<u>ทดสอบพารามิเตอร์ a</u> ซึ่งตั้งสมมติฐาน $H_0: a = a_0$ $H_1: a \neq a_0$

ตัวสถิติที่ใช้ทดสอบ คือ

$$t = \frac{a - a_0}{S\sqrt{\sum X^2}/\sqrt{n\sum(X - \bar{X})^2}}$$

เมื่อ ค่า S ได้จาก $\sqrt{MSE}$

ถ้า $|t_c| > t_{1-\alpha/2,n-2}$ เราจะปฏิเสธ H_0

7.1.3.2 การหาช่วงความเชื่อมั่น

ก. ค่าพารามิเตอร์ b

$100(1-\alpha)$ % ช่วงความเชื่อมั่นสำหรับ b คือ

$$\hat{b} \pm t_{1-\alpha/2,n-2} \sqrt{\frac{MSE}{\sum(X - \bar{X})^2}}$$

ข. ค่าพารามิเตอร์ a

100 (1 - α) % ช่วงความเชื่อมั่นสำหรับ a คือ

$$\hat{a} \pm t_{1-\alpha/2,n-2} \sqrt{\frac{MSE(\sum X^2)}{n\sum(X-\bar{X})^2}}$$

ค. ค่าเฉลี่ยของ Y_i (μ_Y) เมื่อ X = X_i และค่า Y_i เมื่อ X = X_i

100 (1 - α) % ช่วงความเชื่อมั่นสำหรับ μ_Y คือ

$$\hat{Y}_i \pm t_{1-\alpha/2,n-2} \sqrt{MSE[\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}]}$$

100 (1 - α) % ช่วงความเชื่อมั่นสำหรับ Y_i คือ

$$\hat{Y}_i \pm t_{1-\alpha/2,n-2} \sqrt{MSE\{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}\}}$$

ตัวอย่างที่ 7.3 จากข้อมูลต่อไปนี้ใน regression model $Y = \beta_0 + \beta_1 X + \in$

1. จงคำนวณหาเส้นสมการถดถอย
2. จงทดสอบสมมติฐานเกี่ยวกับ β_1 โดยใช้ F - test
3. จงทดสอบสมมติฐานเกี่ยวกับ β_0 , β_1 โดยใช้ t - test
4. จงหา 95 % ช่วงความเชื่อมั่นสำหรับ β_0 และ จงหา 95 % ช่วงความเชื่อมั่นสำหรับ β_1
5. จงหา 95 % ช่วงความเชื่อมั่นสำหรับ μ_Y เมื่อ X = 31 และ จงหา 95 % ช่วงความเชื่อมั่นสำหรับ Y_i เมื่อ X = 31

X	Y	XY	X^2	Y^2	$\hat{Y}_i$	Y - $\hat{Y}_i$	(Y- $\hat{Y}_i$)^2
31	35	1085	961	1225	36.046	-1.046	1.094
19	36	684	361	1296	26.578	9.422	88.774
37	32	1184	1369	1024	40.780	-8.780	88.088
57	50	2850	3249	2500	56.560	-6.560	43.034
54	69	3726	2916	4761	54.193	14.807	219.247
51	52	2652	2601	2704	51.826	0.174	0.030
33	29	957	1089	841	37.624	-8.624	74.373
58	56	3248	3364	3136	57.349	-1.349	1.820
19	26	494	361	676	26.578	-0.578	0.334
41	40	1640	1681	1600	43.936	-3.936	15.492
43	52	2236	1849	2704	45.514	6.486	42.068
443	477	20756	19801	22467			563.354

คำตอบ 1. จาก $\hat{Y}_i = a + bX$

ได้ $b = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \dfrac{11(20756) - (443)(477)}{11(19801) - (443)^2}$

$= 0.789$

$a = \bar{Y} - b\,\bar{X} = \dfrac{477}{11} - (0.789)\left(\dfrac{443}{11}\right) = 11.589$

เส้นสมการถดถอย คือ $\hat{Y}_i = 11.589 + 0.789\,X$

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

$SST = \sum Y^2 - [(\sum Y)^2 / n] = 22467 - [(477)^2 / 11] = 1782.545$

$SSR = b[\sum XY - \{(\sum X)(\sum Y)/n\}] = (0.789)[20756 - \{(443)(477)/11\}]$

$= (0.789)(1545.909) = 1219.722$

$SSE = SST - SSR = 1,782.545 - 1,219.722 = 562.823$

ANOVA

SOV.	Df	SS	MS	F
Regression	1	1,219.722	1,219.722	19.5
Residual	9	562.823	62.5359	
Total	10	1,782.545		

$F_{0.05,(1,9)} = 5.12 \quad F_c = 19.5$ นั่นคือ $F_c > F_{0.05,(1,9)}$

นั่นคือ เราปฏิเสธ H_0

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

ตัวสถิติที่ใช้ทดสอบ คือ

$$t = \frac{a - a_0}{S\sqrt{\sum X^2}\Big/\sqrt{n\sum(X - \bar{X})^2}}$$

$$\sqrt{\frac{MSE(\sum X^2)}{n\sum(X - \bar{X})^2}} = \sqrt{\frac{(62.5359)(19,801)}{(11)(1,960.182)}} = \sqrt{57.4827}$$

$= 7.5817$

$t = \dfrac{11.589}{7.5817} = 1.53$ เมื่อ $t_{0.025,9} = 2.262$

นั่นคือ $t < t_{0.025,9}$ เรายอมรับ $H_0: \beta_0 = 0$

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

ตัวสถิติที่ใช้ทดสอบ คือ

$$t = \frac{b - b_0}{S/\sqrt{\Sigma(X_i - \bar{X})^2}}$$ เมื่อค่า S ได้จาก $\sqrt{MSE}$

$$\sqrt{\frac{MSE}{\Sigma(X - \bar{X})^2}} = \sqrt{\frac{62.5359}{1,960.182}} = \sqrt{0.0319} = 0.1786$$

$$t = \frac{0.789}{0.1786} = 4.42 \quad \text{และ เมื่อ} \quad t_{0.025,9} = 2.262$$

$\therefore t > t_{0.025,9}$ นั่นคือ ปฏิเสธ $H_0 : \beta_1 = 0$ คือ b มีนัยสำคัญทำให้ค่าของ Y เปลี่ยนไป เมื่อ X เปลี่ยนค่าไป

4. 95 % ช่วงความเชื่อมั่นสำหรับ β_1 คือ

$$\hat{b} \pm t_{0.025,9}\sqrt{\frac{MSE}{\Sigma(X - \bar{X})^2}} = 0.789 \pm (2.262)(0.1786) = (0.385, 1.193)$$

95 % ช่วงความเชื่อมั่นสำหรับ a คือ

$$= \hat{a} \pm t_{0.025,9}\sqrt{\frac{MSE(\Sigma X^2)}{n\Sigma(X - \bar{X})^2}} = 11.589 \pm (2.262)(7.5817)$$

$$= (-5.563, 28.737)$$

5. 95 % ช่วงความเชื่อมั่นสำหรับ μ_Y เมื่อ X = 31 คือ

$$\hat{Y}_i \pm t_{0.025,9}\sqrt{MSE[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}]}$$

$$\hat{Y} = 11.589 + (0.789)X = 11.589 + (0.789)(31) = 36.049$$

$$\sqrt{MSE[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}]} = \sqrt{(62.5359)\{\frac{1}{11} + \frac{(31 - 40.27)^2}{1,960.182}\}}$$

$$= \sqrt{8.42358} = 2.9023$$

95 % ช่วงความเชื่อมั่นสำหรับ μ_Y เมื่อ X = 31 คือ

$36.049 \pm (2.262)(2.9023) = (29.484, 42.614)$

95 % ช่วงความเชื่อมั่นสำหรับ Y_i เมื่อ X = 31 คือ

$$\hat{Y}_i \pm t_{0.025,9}\sqrt{MSE\{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\}}$$

$$= 36.049 \pm (2.262)(8.4237) = (16.995, 22.103)$$

7.2 Multiple Regression

7.2.1 จุดประสงค์ในการวิเคราะห์เกี่ยวกับ Multiple Regression มีดังนี้

- เพื่อหาสมการที่ใช้ในการพยากรณ์ เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ และ ตัวแปรตาม
- เพื่อทดสอบว่า จะพยากรณ์ตัวแปรตามจากตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ โดยใช้ตัวแบบที่กำหนดได้หรือไม่
- เพื่อทดสอบว่า การพิจารณาตัวแปรอิสระว่า ควรจะมีกี่ตัว หรือ พิจารณาว่า ข้อเท็จจริงของตัวแปรตามน้อยลงสักเท่าไร เมื่อเอาตัวแปรอิสระออกไป
- เพื่อพิจารณาว่า จะพยากรณ์ค่าตัวแปรตามจากตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ ได้มากน้อยอย่างไร
- เพื่อทดสอบว่า จะเปลี่ยนแปลงแก้ไขตัวแบบ ที่แสดงความสัมพันธ์ระหว่าง ตัวแปรตาม และตัวแปรอิสระได้หรือไม่ โดยการวิเคราะห์ residual
- เพื่อ จะได้จัดลำดับความสำคัญของตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ โดยพิจารณาจากการทำ regression

7.2.2 การประมาณค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน

จากวิธีการวิเคราะห์การถดถอยอย่างง่ายนั้น Y เป็นตัวแปรตาม ซึ่งสามารถทำนายได้จากตัวแปร X ตัวเดียว แต่ในความเป็นจริงนั้น ตัวแปร X หลาย ๆ ตัว จึงจะใช้ทำนายตัวแปร Y ซึ่งเรียกว่า การวิเคราะห์การถดถอยเชิงพหุ ซึ่งถือว่าเป็นวิธีการทางสถิติที่นิยมนำไปใช้กันมาก ซึ่งมีตัวแบบคือ

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \in$$

เมื่อ Y เป็นตัวแปรตาม หรือ ตัวแปรที่จะทำนาย

$a, b_1, b_2, \ldots, b_k$ เป็นสัมประสิทธิ์ของการถดถอย (partial regression coefficient)

$X_1, X_2, \ldots, X_k$ เป็นตัวแปรอิสระ

$\in$ เป็นค่าพารามิเตอร์

และได้สมการถดถอย คือ $\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

โดยตัวแปรแต่ละตัวเก็บรวบรวมค่าสังเกต n ค่า ดังนี้

Observation	Variable 1	Variable 2	Variable 3 ...	Variable k
1	X_{11}	X_{12}	X_{13} ...	X_{1k}
2	X_{21}	X_{22}	X_{23} ...	X_{2k}
3	X_{31}	X_{32}	X_{33} ...	X_{3k}
4	X_{41}	X_{42}	X_{43} ...	X_{4k}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	X_{n1}	X_{n2}	X_{n3} ...	X_{nk}

หรืออาจเขียนตัวแบบได้ดังนี้

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + \in_i$$

จะได้สมการปกติ ดังนี้

$$\sum_{i=1}^{n} Y_i = n\,a + b_1 \sum_{i=1}^{n} X_{i1} + b_2 \sum_{i=1}^{n} X_{i2} + \ldots + b_k \sum_{i=1}^{n} X_{ik}$$

$$\sum_{i=1}^{n} Y_i X_{i1} = a \sum_{i=1}^{n} X_{i1} + b_1 \sum_{i=1}^{n} X_{i1} X_{i1} + b_2 \sum_{i=1}^{n} X_{i2} X_{i1} + \ldots + b_k \sum_{i=1}^{n} X_{ik} X_{i1}$$

$$\sum_{i=1}^{n} Y_i X_{i2} = a \sum_{i=1}^{n} X_{i2} + b_1 \sum_{i=1}^{n} X_{i1} X_{i2} + b_2 \sum_{i=1}^{n} X_{i2} X_{i2} + \ldots + b_k \sum_{i=1}^{n} X_{ik} X_{i2}$$

.

.

.

$$\sum_{i=1}^{n} Y_i X_{ik} = a \sum_{i=1}^{n} X_{ik} + b_1 \sum_{i=1}^{n} X_{i1} X_{ik} + b_2 \sum_{i=1}^{n} X_{i2} X_{ik} + \ldots + b_k \sum_{i=1}^{n} X_{ik} X_{ik}$$

การแก้สมการปกติเพื่อประมาณค่า $a, b_1, b_2, \ldots, b_k$ เราอาจใช้เมตริกซ์ในการแก้สมการ ได้ดังนี้

จากตัวแบบ $\underline{Y} = \underline{X}\,\underline{B} + \underline{\in}$

เมื่อ

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & . & X_{1k} \\ 1 & X_{21} & X_{22} & . & . & . & X_{2k} \\ 1 & X_{31} & X_{32} & . & . & . & X_{3k} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & X_{n1} & X_{n2} & . & . & . & X_{nk} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix} \qquad \underline{\in} = \begin{bmatrix} \in_1 \\ \in_2 \\ \in_3 \\ . \\ . \\ . \\ \in_n \end{bmatrix}$$

ในการหาค่าของเมตริกซ์ B โดยวิธีกำลังสองน้อยที่สุด ดังนี้

$$\underline{\in}'\underline{\in} = (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$$
$$= \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} + \mathbf{B}'\mathbf{X}'\mathbf{XB}$$
$$= \mathbf{Y}'\mathbf{Y} - 2\,\mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB}$$
$$\frac{\partial(\underline{\in}'\underline{\in})}{\partial \mathbf{B}} = -2\,\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{XB} = \mathbf{0}$$

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

<u>ในกรณีที่ตัวแปรอิสระมี 2 ตัว</u>

ตัวอย่างที่ 7.4 ในการทดลอง ที่จะแสดงว่าน้ำหนักของสัตว์สามารถที่จะประมาณได้จากน้ำหนักเริ่มต้น และ ประมาณอาหารที่รับประทานเข้าไปในช่วงเวลาหนึ่ง ได้ข้อมูลดังนี้

น้ำหนักสุดท้าย	95	77	80	100	97	70	50	80	92	84
น้ำหนักเริ่มต้น	42	33	33	45	39	36	32	41	40	38
อาหาร	272	226	259	292	311	183	173	236	230	235

ได้ n =10 , $\sum_{i=1}^{n} Y_i = 825$, $\sum_{i=1}^{n} X_{i1} = 379$, $\sum_{i=1}^{n} X_{i2} = 2{,}417$, $\sum_{i=1}^{n} X_{i1}^2 = 14{,}533$,

$\sum_{i=1}^{n} X_{i2}^2 = 601{,}365$, $\sum_{i=1}^{n} X_{i1} X_{i2} = 92{,}628$, $\sum_{i=1}^{n} X_{i1} Y_i = 31{,}726$, $\sum_{i=1}^{n} X_{i2} Y_i = 204{,}569$

ตัวแบบของการถดถอย คือ $Y_i = a + b_1 X_{i1} + b_2 X_{i2} + \in_i$

หรือเขียนอยู่ในรูปเมตริกซ์ ได้ดังนี้

$$Y = \begin{bmatrix} 95 \\ 77 \\ 80 \\ . \\ . \\ . \\ 84 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 42 & 272 \\ 1 & 33 & 226 \\ 1 & 33 & 259 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 38 & 235 \end{bmatrix} \qquad B = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & . & . & . & 1 \\ .42 & 33 & 33 & . & . & . & 38 \\ 272 & 226 & 259 & . & . & . & 235 \end{bmatrix} \begin{bmatrix} 1 & 42 & 272 \\ 1 & 33 & 226 \\ 1 & 33 & 259 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 38 & 235 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 10 & 379 & 2417 \\ 379 & 14533 & 92628 \\ 2417 & 92628 & 601365 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & . & . & . & 1 \\ .42 & 33 & 33 & . & . & . & 38 \\ 272 & 226 & 259 & . & . & . & 235 \end{bmatrix} \begin{bmatrix} 95 \\ 77 \\ 80 \\ . \\ . \\ . \\ 84 \end{bmatrix} = \begin{bmatrix} 825 \\ 31726 \\ 204569 \end{bmatrix}$$

$$\underline{B} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} = \begin{bmatrix} -22.9915 \\ 1.3957 \\ 0.2176 \end{bmatrix}$$

ได้สมการถดถอย คือ

$$\hat{Y} = -22.9915 + 1.3957 X_1 + 0.2176 X_2$$

หรือจะหาค่า a , b_1 , b_2 โดยใช้สูตร ดังต่อไปนี้

$$b_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$b_2 = \frac{(\Sigma x_2 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

เมื่อ $\Sigma x_1 x_2 = \Sigma X_1 X_2 - [(\Sigma X_1)(\Sigma X_2) / n] = 92,628 - [(379)(2417)/10] = 1023.7$

$\Sigma x_1 y = \Sigma X_1 Y - [(\Sigma X_1)(\Sigma Y) / n] = 31,726 - [(379)(825)/10] = 458.5$

$\Sigma x_1^2 = \Sigma X_1^2 - [(\Sigma X_1)^2 / n] = 14,533 - [(379)^2 / 10] = 168.9$

$\Sigma x_2^2 = \Sigma X_2^2 - [(\Sigma X_2)^2 / n] = 601,365 - [(2417)^2 / 10] = 17,176.1$

$\Sigma x_2 y = \Sigma X_2 Y - [(\Sigma X_2)(\Sigma Y) / n] = 204,569 - [(2417)(825)/10] = 5166.5$

$$b_1 = \frac{(458.5)(17176.1) - (5166.5)(1023.7)}{(168.9)(17176.1) - (1023.7)^2} = 1.39567$$

$$b_2 = \frac{(5166.5)(168.9) - (458.5)(1023.7)}{(168.9)(17176.1) - (1023.7)^2} = 0.21761$$

$a = \bar{Y} - b_1 \bar{X}_1 - b_1 \bar{X}_2 = 82.5 - (1.3957)(37.9) - (0.2176)(241.7)$

$a = 82.5 - 105.4915 = -22.9915$

และได้สมการถดถอย คือ

$$\hat{Y} = -22.9915 + 1.3957 X_1 + 0.2176 X_2$$

7.2.3 การทดสอบสมมติฐาน

ตั้งสมมติฐาน คือ $H_0 : b_1 = b_2 = b_3 = \ldots = b_k = 0$

$H_1 : b_i \neq b_j$ สำหรับคู่ลำดับ (i , j) อย่างน้อยหนึ่งคู่ เมื่อ $i \neq j$

ตัวสถิติที่ใช้ทดสอบ คือ $F = \frac{MSR}{MSE}$

เมื่อ $SST = \underline{Y}'\underline{Y} - n\,\bar{Y}^2$ $\quad SSR = \underline{B}'\underline{X}'\underline{Y} - n\,\bar{Y}^2$

หรือ $SSR = b_1 \sum x_1 y + b_2 \sum x_2 y$ $\quad SSE = SST - SSR = \underline{Y}'\underline{Y} - \underline{B}'\underline{X}'\underline{Y}$

ANOVA

SOV.	df	Sum of Square	Mean Square	F
Regression (b_1, b_2 / a)	k	SSR	SSR / k	MSR / MSE
Residual	n - k - 1	SSE	SSE / (n - k - 1)	
Total	n - 1	SST		

ถ้า $F_c > F_{\alpha,(k,n-k-1)}$ เราจะปฏิเสธ H_0

ตัวอย่างที่ 7.5 จากข้อมูลในตัวอย่างที่ 7.4 จงทดสอบสมมติฐาน

$H_0 : b_1 = b_2 = 0$ $\quad H_1 : b_1 \neq b_2$

$SST = \sum y^2 = \sum Y^2 - [(\sum Y)^2 / n] = 70{,}083 - [(825)^2 / 10] = 2{,}020.5$

$SSR = (1.3957)(458.5) + (0.2176)(5166.5) = 1{,}764.159$

$SSE = SST - SSR = 2{,}020.5 - 1{,}764.159 = 256.341$

ANOVA

SOV.	df	Sum of Square	Mean Square	F
Regression (b_1, b_2 / a)	2	1764.159	882.0795	24.09
Residual	7	256.341	36.62	
Total	9	2020.5		

ถ้า $F_c > F_{0.05,(2,7)}$ เมื่อ $F_{0.05,(2,7)} = 4.74$ นั่นคือ เราจะปฏิเสธ H_0

7.2.4 การหาช่วงความเชื่อมั่น

$100(1-\alpha)$ % ช่วงความเชื่อมั่นสำหรับ b_j คือ $\hat{b}_j \pm t_{\alpha/2,\ n-k-1}\ S(\hat{b}_j)$

เมื่อ $S^2(\hat{b}_j) = C_{j+1}S^2 = C_{j+1}(MSE)$

C_{j+1} เป็นสัมประสิทธิ์ตัวที่ $j+1$ ซึ่งเป็นสมาชิกในแนวทแยงหลักของเมตริกซ์ $(X'X)^{-1}$

$100(1-\alpha)$ % ช่วงความเชื่อมั่นสำหรับค่าเฉลี่ยของ Y หรือ $E(Y/X_1, X_2, \ldots, X_k)$

คือ $\hat{Y}_p \pm t_{\alpha/2,\ n-k-1}\ S\sqrt{X'_p(X'X)^{-1}X_p}$

$100(1-\alpha)$ % ช่วงความเชื่อมั่นสำหรับค่าของ Y คือ

$\hat{Y}_p \pm t_{\alpha/2,\ n-k-1}\ S\sqrt{1 + [X'_p(X'X)^{-1}X_p]}$

เมื่อ $\hat{Y}_p = X'_p B$

และ X_p เป็นเมตริกซ์ของค่า X ที่โจทย์กำหนดให้โดยเฉพาะ

ตัวอย่างที่ 7.6 จากข้อมูลในตัวอย่างที่ 7.4 จงหา

1. จงทดสอบสมมติฐาน $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$ โดยใช้ t - test

และ จงทดสอบสมมติฐาน $H_0 : \beta_2 = 0$ $H_1 : \beta_2 \neq 0$ โดยใช้ t - test

2. จงหา 95 % ช่วงความเชื่อมั่นสำหรับ β_1 และ β_2

3. จงหา 95 % ช่วงความเชื่อมั่นสำหรับ $E(Y/\ X_1 = 40, X_2 = 245)$

4. จงหา 95 % ช่วงความเชื่อมั่นสำหรับ Y เมื่อ $X_1 = 40, X_2 = 245$

<u>คำตอบ</u> 1. $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

$$t = \frac{\hat{b}_1 - 0}{S(\hat{b}_1)}$$ เมื่อ $S(\hat{b}_1) = \sqrt{C_{j+1}(MSE)} = \sqrt{C_2(MSE)}$

เมื่อ $S(\hat{b}_1) = \sqrt{(0.0092688)(36.611567)} = \sqrt{0.3393452} = 0.5825$

C_2 ได้จากสมาชิกตัวที่ 2 ในแนวทแยงหลักของเมตริกซ์ $(\underline{X}'\underline{X})^{-1}$ คือ

$$(\underline{X}'\underline{X})^{-1} = \begin{bmatrix} 8.617548 & -0.2177683 & -0.010928 \\ (a) & & \\ -0.217768 & 0.0092688 & -0.005524 \\ & (b_1) & \\ -0.0010928 & -0.0005524 & 0.0000894 \\ & & (b_2) \end{bmatrix}$$

$C_2 = 0.0092688$ ได้ $t = 1.3957 / 0.5825 = 2.396$

เปรียบเทียบกับ $t_{0.025,7} = 2.365$ $\therefore t_c > t$ จากตาราง

นั่นคือ ปฏิเสธ . $H_0 : \beta_1 = 0$

$H_0 : \beta_2 = 0$ $H_1 : \beta_2 \neq 0$

$t = \dfrac{\hat{b}_2 - 0}{S(\hat{b}_2)}$ เมื่อ $S(\hat{b}_2) = \sqrt{C_{j+1}(MSE)} = \sqrt{C_3(\text{MSE})}$

เมื่อ $S(\hat{b}_2) = \sqrt{(0.0000894)(36.611567)} = \sqrt{0.003273074} = 0.0572$

$t = 0.2176 / 0.0572 = 3.804$ เมื่อ t จากตาราง คือ $t_{0.025,7} = 2.365$

$\therefore t_c > t$ จากตาราง

นั่นคือ ปฏิเสธ . $H_0 : \beta_2 = 0$

2. 95 % ช่วงความเชื่อมั่นสำหรับ b_1 คือ

$\hat{b}_1 \pm t_{0.025,7} S(\hat{b}_1) = 1.3957 \pm (2.365)(0.5825) = (0.0181, 2.7733)$

95 % ช่วงความเชื่อมั่นสำหรับ b_2 คือ

$\hat{b}_2 \pm t_{0.025,7} S(\hat{b}_2) = 0.2176 \pm (2.365)(0.0572) = (0.0823, 0.3529)$

3. 95 % ช่วงความเชื่อมั่นสำหรับ $E(Y / X_1 = 40, X_2 = 245)$ คือ

$\hat{Y}_p \pm t_{0.025,7} \; S \sqrt{X'_p (X'X)^{-1} X_p}$

เมื่อ $X'_p = [\,1 \quad 40 \quad 245\,]$

$\hat{Y}_p = -22.9915 + 1.3957 X_1 + 0.2176 X_2 = -22.9915 + 1.3957(40) + 0.2176(245)$

$= 86.1485$

$X'_p (X'X)^{-1} X_p =$

$$= [\,1 \quad 40 \quad 245\,] \begin{bmatrix} 8.617548 & -0.2177683 & -0.010928 \\ (a) & & \\ -0.217768 & 0.0092688 & -0.005524 \\ & (b_1) & \\ -0.0010928 & -0.0005524 & 0.0000894 \\ & & (b_2) \end{bmatrix} \begin{bmatrix} 1 \\ 40 \\ 245 \end{bmatrix}$$

$= 21.683967$

$(\text{MSE})[\, X'_p (X'X)^{-1} X_p \,] = (36.611567)(21.683967) = 793.8840$

95 % ช่วงความเชื่อมั่นสำหรับ $E(Y/X_1 = 40, X_2 = 245)$ คือ

$(86.1485) \pm (2.365)\sqrt{793.8840}$

95 % ช่วงความเชื่อมั่นสำหรับค่าของ Y เมื่อ $X_1 = 40, X_2 = 245$ คือ

$$\hat{Y}_p \pm t_{0.025,7}\ s\sqrt{1 + [X'_p (X'X)^{-1} X_p]}$$

$$= 86.1485 \pm (2.365)\sqrt{(36.611567)(1 + 21.683967)} = (17.993, 154.304)$$

7.3 Multiple Correlation and the Coefficient of Determination

ในกรณีของสหสัมพันธ์อย่างง่าย (simple correlation) เมื่อยกกำลังสอง และคำนวณเป็น factor ใหม่ เรียกว่า coefficient of determination หรือ R^2 มีค่าคือ

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad \text{หรือ} \qquad R^2 = \frac{\hat{B}'\underline{X}\,\underline{Y} - n\bar{Y}^2}{\underline{Y}\,\underline{Y} - n\bar{Y}^2}$$

R^2 มีค่าจาก 0 ถึง 1 ซึ่งแทนความผันแปรที่สามารถอธิบายได้เหนือความผันแปรทั้งหมด ซึ่งจากสมการดังกล่าวนี้สามารถใช้ได้ทั้งใน simple regression และ multiple regression หรือค่า R^2 เป็นการบอกว่าสมการถดถอยที่ใช้ในการพยากรณ์สามารถอธิบายตัวแปรตามได้มากน้อยเพียงใด ค่า R เราเรียกว่า multiple correlation coefficient แต่นิยมใช้ค่า R^2 อธิบายความหมายของสมการถดถอยที่ได้มากกว่าค่า R เพราะ ค่า R^2 ให้ความหมายได้ชัดเจนมากกว่า

simple correlation coefficient ของตัวแปร Y กับ X_1 แทนด้วย r_{Y1}

$$r_{Y1} = \frac{\sum yx_1}{\sqrt{(\sum y^2)(\sum x_1^2)}}$$

simple correlation coefficient ของตัวแปร Y กับ X_2 แทนด้วย r_{Y2}

$$r_{Y2} = \frac{\Sigma yx_2}{\sqrt{(\Sigma y^2)(\Sigma x_2^2)}}$$

simple correlation coefficient ของตัวแปร X_1 กับ X_2 แทนด้วย r_{12}

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}}$$

partial correlation coefficient ของตัวแปร Y กับ X_1 เมื่อตัวแปร X_2 อยู่ในสมการแล้ว (หรือถือว่าตัวแปร X_2 คงที่) จะได้

$$r_{Y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

partial correlation coefficient ของตัวแปร Y กับ X_2 เมื่อตัวแปร X_1 อยู่ในสมการแล้ว (หรือถือว่าตัวแปร X_1 คงที่) จะได้

$$r_{Y2.1} = \frac{r_{y2} - r_{y1}.r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}$$

และ $$r_{Y2.3} = \frac{r_{y2} - r_{y3}.r_{23}}{\sqrt{(1-r_{y3}^2)(1-r_{23}^2)}}$$

partial correlation coefficient ของตัวแปร Y กับ X_2 เมื่อตัวแปร X_3 และ X_4 อยู่ในสมการแล้ว (หรือถือว่าตัวแปร X_3 และ X_4 คงที่) จะได้

$$r_{Y2.34} = \frac{r_{y2.4} - r_{y3.4}r_{23.4}}{\sqrt{(1-r_{y3.4}^2)(1-r_{23.4}^2)}}$$

หรืออาจเท่ากับ $$\frac{r_{y2.3} - r_{y4.3}r_{24.3}}{\sqrt{(1-r_{y4.3}^2)(1-r_{24.3}^2)}}$$

เมื่อเราทราบค่า simple correlation coefficient ของตัวแปร Y กับ X_1 (r_{y1}) และ r_{y2} และ r_{12} เราสามารถหาค่า R^2 ได้จาก

$$R^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1-r_{12}^2}$$

ตัวอย่างที่ 7.7 จากข้อมูลในตัวอย่างที่ 7.4 จงหาค่า r_{Y1} , r_{Y2} , r_{12} , R^2 , $r_{Y1.2}$

$$r_{Y1} = \frac{\sum yx_1}{\sqrt{(\sum y^2)(\sum x_1^2)}} = \frac{458.5}{\sqrt{(2020.5)(168.9)}} = 0.7849$$

simple correlation coefficient ของตัวแปร Y กับ X_2 แทนด้วย r_{Y2}

$$r_{Y2} = \frac{\Sigma yx_2}{\sqrt{(\Sigma y^2)(\Sigma x_2^2)}} = \frac{5166.5}{\sqrt{(2020.5)(17176.1)}} = 0.8770$$

simple correlation coefficient ของตัวแปร X_1 กับ X_2 แทนด้วย r_{12}

$$r_{12} = \frac{\Sigma x_1x_2}{\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}} = \frac{1023.1}{\sqrt{(168.9)(17176.1)}} = 0.6007$$

R^2 = SSR / SST = 1764.159 / 2020.5 = 0.873

หรือ $$R^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1-r_{12}^2} = \frac{(.7849)^2 + (.877)^2 - 2(.7849)(.877)(.6007)}{1-(.6007)^2}$$

R^2 = 0.5582 / 0.6392 = 0.873

$$r_{Y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

$$= \frac{(.7849) - (.877)(.6007)}{\sqrt{[1-(.877)^2][1-(.6007)^2]}} = 0.258 / 0.384$$

$$= 0.6719$$

7.4 การเลือกสมการถดถอยที่ดีที่สุด (Selection the best regression)

เราต้องการสร้างสมการของการพยากรณ์ในรูปของสมการการถดถอย สำหรับตัวแปรตาม Y ในเทอมของตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ เราสมมติว่า เซตของตัวแปรจากสมการถูกเลือกและรวมอยู่ในฟังก์ชัน (อาจจะเป็นรูปกำลังสอง และ รูปผลคูณของตัวแปร) เพื่อให้เกิดความพึงพอใจและความจำเป็นที่จะนำไปใช้ ดังนั้น การเลือกสมการถดถอยที่ดีที่สุด ว่าจะรวมตัวแปรในสมการกี่ตัวดี มีเหตุผล 2 เหตุผล ดังนี้

1. สร้างสมการที่มีประโยชน์ให้ตรงกับจุดประสงค์ในการทำนายสิ่งที่เราต้องการตัวแบบที่รวมตัวแปร X ต้องใช้กี่ตัวที่เป็นไปได้ และ ให้ความถูกต้องแม่นยำในการประมาณค่าเพื่อการตัดสินใจ
2. เพราะค่าใช้จ่ายในการรวบรวมข้อมูลข่าวสาร สำหรับจำนวนของตัวแปรอิสระ X ที่มีจำนวนมากย่อมมีค่าสูงกว่าจำนวนของตัวแปรอิสระ X ที่น้อยกว่า

ในกระบวนการทางสถิติสำหรับเลือกสมการถดถอยที่ดีที่สุด มิได้มีเพียงวิธีเดียว การวินิจฉัยส่วนตัวจึงเป็นสิ่งจำเป็นส่วนหนึ่งของวิธีการทางสถิติที่จะต้องมีการอภิปรายกัน ในหัวข้อนี้เราจะอธิบายถึงกระบวนการทั้งหลาย เพื่อให้ไว้พิจารณาวิธีการทั้งหลายที่ปรากฏนี้ ปัจจุบันยังนำไปใช้กันอยู่ เพื่อไม่ให้เกิดความสับสน วิธีการทั้งหลายอาจจะไม่นำไปสู่คำตอบที่เหมือน ๆ กัน เมื่อเราใช้ปัญหาเดียวกัน ถึงแม้ว่าสำหรับหลาย ๆ ปัญหาอาจนำไปสู่การได้รับคำตอบเดียวกันก็ตาม ซึ่งวิธีการทั้งหลาย มีดังนี้

1. All possible regression
2. Backward elimination
3. Forward selection
4. Stepwise regression
5. Two variations on the four previous methods
6. Stagwise regression

7.4.1 All Possible Regression

วิธีการนี้ไม่เหมาะแก่การใช้วิธีหนึ่ง ถ้าปราศจากการปฏิบัติการเกี่ยวกับข้อมูล โดยใช้คอมพิวเตอร์ที่มีความเร็วสูง ๆ ถ้าใช้คอมพิวเตอร์ที่มีความเร็วสูง มีประสิทธิภาพ ในการดำเนินงานทางโปรแกรม จะได้ผลของสิ่งที่ต้องการทราบอย่างมีคุณค่า วิธีการนี้เริ่มจากพิจารณาสมการถดถอยที่เป็นไปได้ทั้งหมดจากตัวแปรอิสระ $X_1, X_2, \ldots, X_k$ ดังนี้

1. สมการถดถอยที่ไม่มีตัวแปรอิสระเลย คือ $\hat{Y} = a$

2. สมการถดถอยที่มีตัวแปรอิสระเพียงตัวเดียว คือ $\hat{Y} = a + b_i X_i \quad ; i = 1,2,\ldots,k$

มีอยู่ ${}^{k}C_1$ สมการ หรือเท่ากับ k สมการ

3. สมการถดถอยที่มีตัวแปรอิสระ 2 ตัว คือ

$$\hat{Y} = a + b_i X_i + b_j X_j \quad ; \quad i = 1,2,\ldots,k \quad j = 1,2,\ldots,k$$

และ $i \neq j$

มีสมการอยู่ ${}^{k}C_2$ สมการ

4. หาสมการที่มีตัวแปรอิสระ 3 ตัวเรื่อย ๆ ไปจนถึงสมการถดถอยที่มีตัวแปรอิสระ k ตัว ซึ่งจะได้สมการถดถอยทั้งหมด 2^k สมการ จากนั้นให้คำนวณค่า R^2 ของทุก ๆ สมการ พิจารณาสมการถดถอยที่มีตัวแปรอิสระเท่ากันในแต่ละกลุ่ม เลือกเฉพาะสมการ ที่ให้ค่า R^2 สูงที่สุด ของแต่ละกลุ่มมาพิจารณาว่าสมการถดถอยใดดีที่สุด โดยดูจากการเพิ่มตัวแปรอิสระเข้าไปในสมการอีก 1 ตัว แล้วค่า R^2 จะเพิ่มมากน้อยเพียงใด และคุ้มกับค่าใช้จ่ายที่เพิ่มขึ้นจากการเก็บรวบรวมข้อมูลเพิ่มขึ้นและค่าใช้จ่ายในการดำเนินงานทั้งสิ้นที่เพิ่มขึ้นหรือไม่อย่างไร วิธีนี้ดีตรงสามารถเปรียบเทียบในแต่ละกลุ่มของสมการ แต่ต้องเสียเวลาในการปฏิบัติงานมาก

7.4.2 Backward Elimination

วิธีนี้ปรับปรุงจากวิธี all possible regression เป็นการพยายามที่จะตรวจสอบสมการถดถอยที่มีตัวแปรอิสระทั้งหมด จนกว่าจะได้สมการถดถอยที่ดีที่สุดที่มีจำนวนตัวแปรอิสระที่แน่นอน ขั้นตอนพื้นฐานของวิธีนี้คือ

1. หาสมการถดถอยที่มีตัวแปรอิสระทั้งหมด คือ

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

2. คำนวณค่า partial F - test ของตัวแปรอิสระทุกตัวที่นำเข้ามาในสมการถดถอยเป็นตัวสุดท้าย

3. เปรียบเทียบค่า partial F - test ตัวที่ต่ำที่สุด สมมติเป็น F_L กับค่า F จากตาราง คือ $F_{\alpha,(1,n-k-1)}$ ให้พิจารณาจาก

3.1 ถ้า F_L < F จากตาราง ให้นำตัวแปร X_L ออกจากสมการถดถอย แล้วหาสมการถดถอยใหม่ที่ไม่รวมตัวแปร X_L แล้วกลับไปทำตามขั้นตอนที่ 2 ใหม่ จนกระทั่งได้ค่า partial F - test ของตัวแปรทุก ๆ ตัวมากกว่า F จากตาราง

3.2 ถ้า F_L > F จากตาราง จะได้สมการถดถอยที่ดีที่สุด วิธีการต่าง ๆ จะหยุดดำเนินการ

ตัวอย่างที่ 7.8 ตัวแปร X_1, X_2, X_3, X_4 เป็นตัวแปรอิสระซึ่งวัดเป็นเปอร์เซนต์ของน้ำหนักของอิฐที่นำมาละลายรวมกันจากเตาจนเป็นซีเมนต์ ส่วนค่า Y หรือ X_5 เป็นความร้อนที่รวมตัวอยู่ในซีเมนต์ หน่วยเป็นแคลอรี่ต่อกรัม จงหาสมการถดถอยที่ดีที่สุด โดยวิธี

1. all possible regression 2. Backward elimination

เมื่อตัวแปรแต่ละตัว คือ

X_1 = ปริมาณ tricalcium aluminate

X_2 = ปริมาณ tricalcium silicate

X_3 = ปริมาณ tetracalcium alumino ferrite

X_4 = ปริมาณ dicalcium silicate

Response = Y = X_5 = ปริมาณความร้อนของซีเมนต์

ข้อมูลเดิม

	X_1	X_2	X_3	X_4	X_5
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4
ค่าเฉลี่ยของแต่ละตัวแปร					
1	7.4615383	48.153845	11.76923	29.999999	95.423075
ค่าส่วนเบี่ยงเบนมาตรฐานของแต่ละตัวแปร					
1	5.8823944	15.560879	6.40512599	16.738178	15.043724
<u>เมตริกซ์ของสัมประสิทธิ์สหสัมพันธ์</u>					
1	.99999991	.22857948	-.82413372	-.24544512	.73071745
2	.22857948	1.0000010	-.13924238	-.97295516	.81625268
3	-.82413372	-.13924238	.99999991	.02953700	-.53467065
4	-.24544512	-.97295516	0.2953700	1.00000010	-.82130513
5	.73071745	.81625268	-.53467065	-.82130513	.99999999

1. all possible regression

การพิจารณาสมการถดถอยที่เป็นไปได้ทั้งหมด

จำนวนตัวแปรอิสระ	สมการถดถอย	R^2 (%)
0	$\hat{Y} = 95.5$	
1	$\hat{Y} = 81.479 + 1.869 X_1$	53.395
	$\hat{Y} = 57.424 + 0.789 X_2$	66.627
	$\hat{Y} = 110.203 - 1.256 X_3$	28.587
	$\hat{Y} = 117.568 - 0.738 X_4$	67.454
2	$\hat{Y} = 52.577 + 1.468 X_1 + 0.662 X_2$	97.868
	$\hat{Y} = 72.349 + 2.312 X_1 + 0.494 X_3$	54.817
	$\hat{Y} = 103.097 + 1.440 X_1 - 0.614 X_4$	97.247
	$\hat{Y} = 72.075 + 0.731 X_2 - 1.008 X_3$	84.703
	$\hat{Y} = 94.160 + 0.311 X_2 - 0.467 X_4$	68.006
	$\hat{Y} = 131.282 - 1.200 X_3 - 0.725 X_4$	93.529
3	$\hat{Y} = 48.194 + 1.696 X_1 + 0.657 X_2 + 0.250 X_3$	98.228
	$\hat{Y} = 71.648 + 1.452 X_1 + 0.416 X_2 - 0.237 X_4$	98.234
	$\hat{Y} = 111.684 + 1.051 X_1 - 0.410 X_3 - 0.643 X_4$	98.128
	$\hat{Y} = 203.642 - 0.923 X_2 - 1.448 X_3 - 1.557 X_4$	98.282
4	$\hat{Y} = 64.405 + 1.551 X_1 + 0.510 X_2 + 0.102 X_3 - 0.144 X_4$	98.238

ประเภทของตัวแปร	ตัวแปรในสมการ	R^2
ตัวเดียว	$\hat{Y} = 117.568 - 0.738 X_4$	67.5 %
สองตัว	$\hat{Y} = 52.577 + 1.468 X_1 + 0.662 X_2$	97.9 %
	$\hat{Y} = 103.097 + 1.440 X_1 - 0.614 X_4$	97.2 %
สามตัว	$\hat{Y} = 71.648 + 1.452 X_1 + 0.416 X_2 - 0.237 X_4$	98.234 %
สี่ตัว	$\hat{Y} = 64.405 + 1.551 X_1 + 0.510 X_2 + 0.102 X_3 - 0.144 X_4$	98.238

พิจารณาจากประเภทของตัวแปรอิสระ 2 ตัว จะมีค่า R^2 ไม่แตกต่างกันมากนัก เมื่อพิจารณาจากเมตริกซ์ของสัมประสิทธิ์สหสัมพันธ์จากข้อมูลที่มีค่าสูง คือ $r_{13} = -0.82413372$ และ $r_{24} =$

- 0.97295516 ดังนั้น เมื่อตัวแปร X_1 และ X_2 หรือเมื่อ X_1 และ X_4 อยู่ในสมการถดถอย จะมีค่าของความแปรปรวนที่ไม่สามารถอธิบายได้ น้อยมาก ต่อค่าตัวแปรตาม จากค่า R^2 ของประเภทสามตัวแปรและสี่ตัวแปร จะได้รับประโยชน์น้อยมาก เมื่อเพิ่มตัวแปรเข้าไปอีก 1 ตัวหรือสองตัว

เมื่อสมการได้ถูกเลือกจากหนึ่งในสองสมการของประเภทสองตัวแปร แต่จะเลือกสมการใด ถ้าเลือก $f(X_1, X_2)$ ซึ่งจะค้านกับกรณีประเภทตัวแปรตัวเดียว ที่บอกว่าสมการถดถอยที่เหมาะสมคือ $\hat{Y} = f(X_4)$ จากเหตุผลดังกล่าวนี้ ผู้พยากรณ์ชอบที่จะเลือก $f(X_1, X_4)$ มากกว่า จะได้สมการถดถอยคือ $\hat{Y} = 103.097 + 1.44\ X_1 - 0.614\ X_4$

2. backward elimination

เริ่มจากหาสมการถดถอยโดยใช้ least square method ได้แก่ $\hat{Y} = f(X_1, X_2, X_3, X_4)$ จากค่าที่ให้มาข้างล่างนี้ จะได้ว่า กระบวนการวิเคราะห์การถดถอยที่จะนำตัวแปรเข้ามาในสมการถดถอย ตามลำดับคือ X_4 ลำดับแรก ต่อมาก็ X_3 และ X_2 และตัวแปรสุดท้าย คือ X_1 เพื่อที่จะกำจัดตัวแปรแต่ละตัวของ X_1, X_2, X_3 และ X_4 ออกจากสมการถดถอย โดยพิจารณาจาก partial F - test ที่อยู่ในคอลัมน์สุดท้าย จะได้ว่า ค่าที่น้อยที่สุด คือ ค่า F - test ของ X_3 ซึ่งเท่ากับ 0.018 แต่ F จากตาราง คือ $F_{0.90,(1,8)} = 3.46$ ซึ่งมีค่าน้อยกว่า F จากตาราง ดังนั้น จึงควรนำ X_3 ออกจากสมการถดถอย และได้ $\hat{Y} = f(X_1, X_2, X_4)$ ได้ค่า F ของ overall ของทั้งสามตัวแปรเท่ากับ 166.84 เปรียบเทียบกับ $F_{0.999,(3,9)} = 13.90$ ได้ว่า F จากการคำนวณ มากกว่า F จากตาราง

ANOVA

SOV.	d.f.	S S	M S	F
Regression	4	2667.9	666.975	111.4795
Residual	8	47.8635	5.983	
Total	12	2715.7635		

ตัวแปร	Partial F - test
1	4.338
2	0.497
3	0.018
4	0.041

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	2667.79	889.26	166.84
Residual	9	47.97	5.33	
Total	12	2715.76		

ตัวแปร	Partial F - test
1	154.008
2	5.026
4	1.863

ตัวแปร	Square of Partials
3	0.002
Y	1.000

แต่ตรวจสอบจากสมการนี้ พิจารณาค่า partial F - test ของตัวแปรแต่ละตัวจะได้ว่า partial F ของตัวแปร X_4 มีค่าน้อยที่สุด คือเท่ากับ 1.86 ซึ่งน้อยกว่า F จากตาราง = $F_{0.90,(1,9)} = 3.36$ นั่นคือ สมการถดถอยที่เหมาะสมที่สุด คือ $\hat{Y} = f(X_1, X_2)$

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	2657.86	1328.93	229.52
Residual	10	57.90	5.79	
Total	12	2715.76		

ตัวแปร	Partial F - test
1	146.52
2	208.58

ตัวแปร	Square of Partials
3	0.169
4	0.172
5	1.000

จากค่า F ของ Regression = 229.5 ซึ่งมีค่ามากกว่า $F_{0.999,(2,10)} = 14.91$ อย่างมีนัยสำคัญ นอกจากนี้พิจารณาค่า partial F ของตัวแปร X_2 และ X_1 ได้ค่า 208.58 และ 146.52 ตามลำดับ ซึ่งมีค่ามากกว่า F จากตาราง จึงได้สมการถดถอยที่ดีที่สุดคือ $\hat{Y} = 52.58 + 0.66 X_2 + 1.47 X_1$

7.4.3 The Forward Selection Procedure

วิธี backward elimination เริ่มจากสมการถดถอยที่โต โดยใช้ตัวแปรทั้งหมด และค่อย ๆ ลดจำนวนตัวแปรในสมการ จนกระทั่งได้สมการถดถอยที่สามารถนำไปใช้ แต่วิธีนี้มีการดำเนินการในทิศทางตรงข้าม โดยสอดแทรกตัวแปรเข้าไปในสมการถดถอย จนกระทั่งเป็นที่พอใจ ลำดับในการนำตัวแปรเข้าไปในสมการถดถอย โดยใช้ partial correlation coefficient เป็นค่าวัดที่สำคัญของตัวแปรว่าจะนำเข้าไว้ในสมการหรือไม่ มีขั้นตอนดังนี้

1. เลือกตัวแปรอิสระ X ที่มีความสัมพันธ์กับ Y มากที่สุด สมมติเป็นตัวแปร X_1 นั่นคือ r_{y1} มีค่ามากที่สุด เราได้ตัวแบบ คือ $\hat{Y} = f(X_1)$

2. จากนั้นคำนวณค่า partial correlation coefficient ของ X_j ($j \neq 1$) และ Y หลังจากที่ตัวแปร X_1 ได้เข้ามาในสมการแล้ว ($r_{yj.1}$) หรือ หมายถึง เราหาความสัมพันธ์ระหว่างค่าความคลาดเคลื่อนจากสมการถดถอยของ $\hat{Y} = f(X_1)$ และค่าความคลาดเคลื่อนจากสมการถดถอยของ $\hat{Y} = f(X_1, X_j)$ ซึ่ง X_j มีค่า partial correlation coefficient กับ Y สูงที่สุดที่ได้เลือก X_j เข้ามาไว้ในสมการ สมมติว่าเป็น X_2 สมการถดถอยที่ได้ คือ $\hat{Y} = f(X_1, X_2)$

3. ดำเนินการตามกระบวนการข้อ 2 โดยพิจารณาค่า partial correlation coefficient กับ Y ที่สูงที่สุดตัวถัดมา สมมติเป็น X_1 เมื่อตัวแปร X_1 และ X_2 ได้เข้ามาในสมการแล้ว (ในที่นี้คือ ค่า $r_{y1.12}$) จะนำตัวแปรตัวนั้นเข้ามา โดยตัวแปรแต่ละตัวที่จะนำเข้ามาในสมการถดถอย ค่าที่จะตรวจสอบมีดังนี้

- ค่า R^2 ของสมการถดถอยเชิงพหุ
- ค่า partial F - test สำหรับตัวแปรส่วนใหญ่ที่นำเข้ามา ซึ่งจะแสดงว่า ตัวแปรที่เพิ่งนำเข้ามาใหม่นั้น มีนัยสำคัญหรือไม่ ถ้าค่า partial F - test มีนัยสำคัญ ให้นำตัวแปรตัวนั้นเข้ามาไว้ในสมการถดถอย แต่ถ้าไม่มีนัยสำคัญ เราจะหยุดดำเนินการ

ตัวอย่างที่ 7.9 จากข้อมูลในตัวอย่างที่ 7.8 จงหาสมการถดถอยที่ดีที่สุด โดยวิธี forward selection

<u>ขั้นตอนที่ 1</u> หาค่า $r_{y1}, r_{y2}, r_{y3}, r_{y4}$ ได้ค่า $r_{y4} = -0.821$ ซึ่งมีค่าสูงที่สุด ดังนั้นให้นำตัวแปร X_4 เข้ามาไว้ในสมการถดถอย เป็นตัวแปรแรก

<u>ขั้นตอนที่ 2</u> จาก overall F - test ของ regression มีนัยสำคัญ ให้นำตัวแปร X_4 เข้ามาไว้ในสมการ ได้ $\hat{Y} = 117.568 - 0.738 X_4$

Variable Entering 4

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	1831.897	1831.897	22.798
Residual	11	883.867	80.352	
Total	12	2715.764		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	- 0.738	0.155	22.799

Constant term 117.568

ตัวแปร	Square of Partials
1	0.915
2	0.019
3	0.801

<u>ขั้นตอนที่ 3</u> คำนวณค่า partial correlation coefficient ของตัวแปรทุกตัวที่ไม่อยู่ในสมการถดถอย กับตัวแปร Y เลือกตัวแปรที่ให้ค่า partial correlation coefficient ที่สูงที่สุด ในที่นี้ คือ ตัวแปร X_1 ซึ่งได้ค่า $r^2_{y1.4} = 0.915$

<u>ขั้นตอนที่ 4</u> เมื่อตัวแปร X_1 และ X_4 อยู่ในสมการถดถอยแล้ว คือ

$\hat{Y} = 103.097 + 1.440 X_1 - 0.614 X_4$

ได้ค่า $R^2 = 97.2\ \%$ ซึ่งมี $F = 176.63$ มากกว่า $F_{0.999,(2,10)} = 14.91$ นั่นคือ ตัวแปร X_1 และ X_4 มีผลต่อตัวแปร Y เมื่อพิจารณาตัวแปร X_1 ให้ค่าผลรวมกำลังสอง ของความคลาดเคลื่อนลดลงอย่างมีนัยสำคัญ โดยดูจากค่า partial F ได้ $= 108.22$ ซึ่งมากกว่า $F_{0.999,(1,10)} = 21.04$

ขั้นตอนที่ 5 ค่า partial correlation coefficient ของตัวแปรที่ยังไม่เข้ามาในสมการ เมื่อยกกำลังสองค่าจะอยู่ช่วงล่าง จะได้ว่า X_2 เป็นตัวแปรตัวต่อมาที่จะนำเข้ามาไว้ในสมการ ได้สมการใหม่ คือ $\hat{Y} = f(X_4, X_1, X_2)$

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	2641.002	1320.501	176.63
Residual	10	74.762	7.476	
Total	12	2715.764		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	1.440	0.138	108.224
4	- 0.614	0.048	159.295

ตัวแปร	Square of Partials
2	0.358
3	0.320

ขั้นตอนที่ 6 สมการถดถอย คือ $\hat{Y} = 71.648 - 0.237 X_4 + 1.452 X_1 + 0.416 X_2$ มีค่า R^2 เพิ่มขึ้นจาก 97.2 % เป็น 98.2 % เมื่อนำ X_2 เข้ามาไว้ในสมการถดถอย มีค่า partial F - test เท่ากับ 5.03 ซึ่งมากกว่า $F_{0.90,(1,9)} = 3.36$ จึงสรุปได้ว่าให้นำตัวแปร X_2 เข้าไว้ในสมการ

จำนวนตัวแปรที่อยู่ในสมการ	3
R^2	98.2

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	2667.79	889.26	166.84
Residual	9	47.97	5.33	
Total	12	2715.76		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
2	0.416	0.186	5.03
1	1.452	0.117	154.01
4	0.237	0.173	1.86

ตัวแปร	Square of Partials
3	0.002

<u>ขั้นตอนที่ 7</u> เมื่อตัวแปรแต่ละตัวที่เข้ามาในสมการอย่างมีนัยสำคัญ ในการทำให้ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองลดลง จากนั้นให้นำ X_3 เข้ามาในสมการ โดยมีค่า partial F เท่ากับ 0.018 ซึ่งไม่มีนัยสำคัญ

ANOVA

SOV.	d.f.	S S	M S	F
Regression	4	2667.9	666.975	111.53
Residual	8	47.86	5.98	
Total	12	2715.76		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	- 0.144	0.709	0.041
3	0.102	0.755	0.018
2	0.510	0.724	0.497
1	1.551	0.745	4.338

จากการวิเคราะห์ความแปรปรวนที่สมบูรณ์ครบถ้วนในทุก ๆ ขั้นตอนได้ผลลัพธ์ ดังนี้

SOV.	d.f.	SS	MS
Total	12	2715.76	
Due to Regression	4	2667.90	
due to X_4	1	1831.90	1831.90
due to $X_1 \mid X_4$	1	809.10	809.10
due to $X_2 \mid X_4, X_1$	1	26.79	26.79
due to $X_3 \mid X_4, X_1, X_2$	1	0.11	0.11
Due to error	8	47.86	5.98

จากวิธีนี้ได้สมการถดถอยที่ดีที่สุด คือ

$$\hat{Y} = 71.648 - 0.237 X_4 + 1.452 X_1 + 0.416 X_2$$

วิธี forward selection เป็นวิธีพื้นฐานที่มีแนวความคิดที่ดี ซึ่งมีสิ่งอำนวยความสะดวก ทางคอมพิวเตอร์มากกว่าวิธีที่กล่าวมาแล้วทั้งสองวิธี ดังนั้น วิธีนี้จึงหลีกเลี่ยงตัวแปร X ที่มีจำนวนมากเกินความจำเป็น ขณะที่แต่ละขั้นตอนจะมีการปรับปรุงสมการถดถอย

7.4.4 The Stepwise Regression Procedure

วิธีการนี้ปรับปรุงมาจากวิธี forward selection โดยปรับปรุงจากการตรวจสอบอีกครั้งในแต่ละขั้นตอนของตัวแปรอิสระที่จะนำเข้ามาในสมการถดถอย ตัวแปรซึ่งอาจจะเป็นตัวแปรตัวเดียวที่ดีที่สุดที่จะนำเข้ามาในสมการถดถอยในขั้นตอนแรก ๆ หรืออาจจะเป็นขั้นตอนต่อมา ซึ่งตัวแปรที่นำเข้ามาอาจจะมากเกินความต้องการ เพราะว่าระหว่างตัวแปรมีความสัมพันธ์กันในตัวแบบถดถอย เพื่อตรวจสอบความสัมพันธ์ดังกล่าวนี้ ให้ใช้ partial F ตรวจสอบแต่ละตัวแปรของสมการถดถอยในแต่ละขั้นตอนของการคำนวณ เพื่อประเมินค่า และเปรียบเทียบด้วยการเลือกจุด ที่เป็นเปอร์เซนต์ของค่า F ที่เหมาะสม สิ่งนี้ทำไว้เพื่อแสดงความคิดเห็นบนการกระจายที่เกิดจากตัวแปรแต่ละตัว ถึงแม้ว่า มีตัวแปรส่วนใหญ่เพิ่งนำเข้ามาตามลำดับในตัวแบบก็ตาม ตัวแปรใดไม่มีนัยสำคัญในการกระจาย เราจะเอาออกจากตัวแบบ และวิธีนี้จะดำเนินไปเรื่อย ๆ จนกว่ามีตัวแปรในสมการได้รับการยอมรับโดยจะปฏิเสธสมมติฐานอย่างมีนัยสำคัญ ดังนี้

1. เริ่มจากหาเมตริกซ์ของความสัมพันธ์อย่างง่ายของตัวแปร Y กับตัวแปรอิสระแต่ละตัว พิจารณาจากความสัมพันธ์ของตัวแปรอิสระที่มีสัมประสิทธิ์สหสัมพันธ์สูงที่สุด สมมติเป็นตัวแปร X_4
2. พิจารณาค่า partial correlation coefficient ที่สูงที่สุด เมื่อตัวแปร X_4 อยู่ในสมการแล้ว ในกรณีตัวอย่างนี้ คือ ตัวแปร X_1
3. ได้สมการถดถอยอยู่ในรูป $\hat{Y} = f(X_4, X_1)$ จากนั้น ให้พิจารณาค่า partial F ของกระบวนการที่ X_1 เข้ามาในสมการก่อน จากนั้นจึงนำ X_4 เข้ามาในสมการเป็นลำดับถัดมา

(ซึ่งวิธี forward selection ไม่ได้ทำกระบวนการนี้) ได้ค่า partial F = 159.295 ซึ่งมีนัยสำคัญที่ระดับ 0.05 ดังนั้น เราจะเก็บตัวแปร X_4 ไว้ในสมการ จากนั้นพิจารณาค่า partial correlation coefficient ที่สูงที่สุด คือ ตัวแปร X_2 ซึ่ง $r^2_{y2.41} = 0.358$

4. สมการถดถอยอยู่ในรูป $\hat{Y} = f(X_4, X_1, X_2)$ ค่า sequential F = 5.026 ซึ่งมากกว่า $F_{0.90,(1,9)} = 3.36$ จึงถือว่า X_2 เข้ามาในสมการอย่างมีนัยสำคัญ จากจุดนี้ ค่า partial F - test ของตัวแปร X_1 และ X_4 นำมาใช้ในการตัดสินใจว่า ยังคงเอาไว้ในสมการถดถอย หรือไม่ จะได้ค่า partial F ของ X_4 มีค่า 1.863 ซึ่งน้อยกว่า $F_{0.90,(1,9)} = 3.36$ จึงควรนำ X_4 ออกจากสมการ

5. ตัวแปรที่ยังเหลืออยู่คือ X_3 ซึ่งค่า sequential F น้อยกว่า F จากตาราง เราจึงปฏิเสธตัวแปร X_3 เราจะได้ว่าสมการถดถอยที่ดีที่สุดจะอยู่ในรูปสมการคือ $\hat{Y} = f(X_1, X_2)$ คือ

$$\hat{Y} = 52.58 + 1.47 X_1 + 0.66 X_2$$

สรุปได้ว่า เราเชื่อว่าวิธีนี้เป็นกระบวนการเลือกตัวแปรที่ดีที่สุด ที่ใช้ในการอภิปรายและแสดงความคิดเห็นในการนำไปใช้ อย่างไรก็ตาม วิธีของ stepwise regression สามารถนำไปใช้ในทางที่ผิดได้สำหรับนักสถิติมือสมัครเล่น ขณะที่กระบวนการอภิปรายทั้งหมดตลอดจนถึงคำวิจารณ์ที่มีเหตุมีผล ยังคงต้องการเกี่ยวกับการเลือกตัวแปรเข้ามาเป็นตัวแรก และในจุดวิกฤติของการตรวจสอบของตัวแบบ จนถึงการตรวจสอบค่าความคลาดเคลื่อนมันเป็นการง่ายที่จะปรับปรุงในการเลือกตัวแปรโดยอัตโนมัติในคอมพิวเตอร์ได้

ตัวอย่างที่ 7.10 แสดงการหาสมการถดถอยที่ดีที่สุดโดยวิธี stepwise regression

ข้อมูลเดิม

ลำดับ	Y	X_1	X_2	X_3	X_4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34

ลำดับ	Y	X_1	X_2	X_3	X_4
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Correlation Matrix ที่ได้ดังนี้

ตัวแปร

1	1.0	0.2286	- 0.8241	- 0.2454	0.7307
2	0.2286	1.0	- 0.1392	- 0.9729	0.8163
3	- 0.8241	- 0.1392	1.0	0.0295	- 0.5347
4	- 0.2454	- 0.9729	0.0295	1.0	- 0.8213
Y	0.7307	0.8163	- 0.5347	- 0.8213	1.0

ขั้นตอนที่1

Variable Entering 4

R - SQ 67.45

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	1831.897	1831.897	22.798
Residual	11	883.867	80.352	
Total	12	2715.764		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	- 0.738	0.155	22.799

ตัวแปร	Square of Partials
1	0.915
2	0.019
3	0.801
Y	1.000

ขั้นตอนที่2

Variable Entering 1

R - SQ 97.247

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	2641.0	1320.5	176.63
Residual	10	74.76	7.476	
Total	12	2715.76		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	- 0.614	0.049	159.295
1	1.440	0.138	108.224

ตัวแปร	Square of Partials
2	0.358
3	0.320
Y	1.000

ขั้นตอนที่3

Variable Entering 2

R - SQ 98.23

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	2667.79	889.26	166.84
Residual	9	47.97	5.33	
Total	12	2715.76		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	- 0.2365	0.173	1.86
1	1.4519	0.117	154.01
2	0.4161	0.186	5.03

ตัวแปร	Square of Partials
3	0.002
Y	1.000

ขั้นตอนที่4

Variable Leaving 4

R - SQ 97.87

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	2657.86	1328.93	229.5
Residual	10	57.90	5.79	
Total	12	2715.76		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	1.468	0.121	146.52
2	0.662	0.046	208.58

ตัวแปร	Square of Partials
3	0.169
4	0.171
Y	1.000

ได้สมการถดถอยที่ดีที่สุดคือ

$$\hat{Y} = 52.58 + 1.47 X_1 + 0.66 X_2$$

ได้ค่าพยากรณ์ของตัวแปรตามดังนี้

ลำดับ	Y	F_t	e_t
1	78.5	80.1	-1.6
2	74.3	73.3	1.0
3	104.3	105.8	-1.5
4	87.6	89.3	-1.7
5	95.9	97.3	-1.4
6	109.2	105.2	4.0
7	102.7	104.0	-1.3
8	72.5	74.6	-2.1
9	93.1	91.3	1.8
10	115.9	114.5	1.4
11	83.8	80.5	3.3
12	113.3	112.4	0.9
13	109.4	112.3	-2.9

7.4.5 Two Variation on the Four Previous Methods

ขณะที่กระบวนการต่าง ๆ ทั้ง 4 วิธี ที่กล่าวมาแล้วไม่สามารถที่จะใช้วิธีใดวิธีหนึ่งที่ดีที่สุดได้ ซึ่งเราอาจจะต้องเลือกตัวแบบที่สามารถยอมรับได้จากวิธีใดวิธีหนึ่ง แต่คงจะไม่เป็นตัวแบบที่ดีที่สุดก็ได้ ดังนั้นทางเลือกใหม่ คือ ให้ประสานวิธีการ 2 วิธีการใด ๆ เมื่อปรับปรุงตัวแบบขึ้นมาให้เป็นตัวแบบที่ดีที่สุด โดยได้เสนอความคิดเห็นไว้ 2 ประการดังนี้

1. ดำเนินตามวิธี stepwise regression ด้วยระดับของการยอมรับและปฏิเสธ เมื่อกระบวนการเลือกสมการหยุดลง ให้ตัดสินใจจำนวนของตัวแปรในขั้นตอนสุดท้ายของการเลือกตัวแบบโดยใช้ค่า R^2 เป็นเกณฑ์ในการตัดสินใจว่าจะใช้จำนวนตัวแปรกี่ตัวจึงจะเหมาะสม เหมือนในวิธี all possible regression โดยจากจำนวนตัวแปรอิสระ k ตัว วิธีนี้จะแก้ปัญหาได้ ในกรณีที่มีรายละเอียดเกี่ยวกับข้อมูลที่จะศึกษาไม่เพียงพอ หรือมีจำนวนข้อมูลไม่เพียงพอไม่สามารถที่จะใช้วิธีการพยากรณ์ วิธีใดวิธีหนึ่งได้ การพิจารณาก่อนหน้านี้ และความคิดเห็นของผู้พยากรณ์ยังคงต้องการที่จะเลือกตัวแบบสุดท้ายของการพยากรณ์ วิธีการนี้เกิดความล้มเหลว เมื่อเซตของตัวแปรโตเกิน แต่ไม่ตรวจสอบโดยใช้วิธี stepwise ในประสบการณ์ของเรา เราจะรวมความได้เปรียบ และ ประโยชน์ของวิธีการนี้เป็นเรื่องรองลงมา สิ่งที่ต้องการมากที่สุด คือการคำนวณทางคอมพิวเตอร์ เพื่อการพยากรณ์

2. เลือกใช้วิธี stepwise regression ด้วยการจำกัดการยอมรับที่น้อยลง และลดระดับนัยสำคัญลง ซึ่งจะส่งผลให้โปรแกรมที่จะยอมรับตัวแปรหลาย ๆ ตัวเข้าไปภายใต้ที่ว่า อะไรจะยอมรับด้วยระดับความไม่เปลี่ยนแปลงที่น้อยลงเหล่านี้ จะยอมให้การตรวจสอบ ที่จะรวมตัวแปรเข้าไปภายใต้กระบวนการของ stepwise โดยปกติและนำไปสู่ตัวแบบที่แตกต่างกันได้

7.4.6 The Stagewise Regression Procedure

วิธีการนี้ไม่ได้ให้ค่าในการแก้สมการ เพื่อประมาณค่าพารามิเตอร์ โดยวิธีกำลังสองน้อยที่สุด ที่ถูกต้องสำหรับตัวแปรที่รวมอยู่ในสมการสุดท้าย แนวความคิดพื้นฐานที่ว่านี้หลังจากสมการถดถอย ที่ตัวแปร X มีความสัมพันธ์มากที่สุดกับตัวแปร Y และ ค่าความคลาดเคลื่อน $Y_i - \hat{Y}_i$ ถูกค้นพบ ค่าความคลาดเคลื่อนนี้ พิจารณาคล้ายกับเป็นค่าของผลกระทบและถดถอยบน X ที่มีความสัมพันธ์มากที่สุดกับตัวแปรตามใหม่นี้ กระบวนการจะดำเนินการไปเรื่อย ๆ ในขั้นตอนต่อ ๆ ไป ซึ่งในแต่ละขั้นตอนจะได้

Response = Fitted Response + (Response - Fitted Response)

สมการถดถอยสามารถใช้แทนกันได้ในขั้นตอนย้อนหลังทีละขั้นตอน จนกระทั่งถึงขั้นตอนสุดท้ายโดยวิธี stagewise จึงจะได้สมการถดถอยโดยไม่ใช้การแก้สมการวิธีกำลังสองน้อยที่สุดในการรวมตัวแปรเข้ามาในสมการ การคำนวณสามารถทำได้ง่ายในการปรับปรุงสมการโดยใช้มือได้ไม่ต้องอาศัยคอมพิวเตอร์เข้ามาช่วยได้เลย

ตัวอย่างที่ 7.11 จากข้อมูลชุดเดิม เราสามารถดำเนินการตามขั้นตอนของวิธี stagwise regression

<u>ขั้นตอนที่ 1</u> ลงชุดค่าตัวแปรตาม Y กับตัวแปรอิสระ X แต่ละตัว ซึ่งจะสอดคล้องกับการตรวจสอบเมตริกซ์ของความสัมพันธ์ที่จะเลือกตัวแปร X ที่มีความสัมพันธ์กับ Y สูงที่สุด

<u>ขั้นตอนที่ 2</u> ได้สมการถดถอยของ Y บน X_4 คือ $\hat{Y} = 117.57 - 0.74 X_4$ และประมาณค่าความคลาดเคลื่อน $Z_i = Y_i - \hat{Y}_i$ ของค่า X_4 แต่ละค่าดังนี้

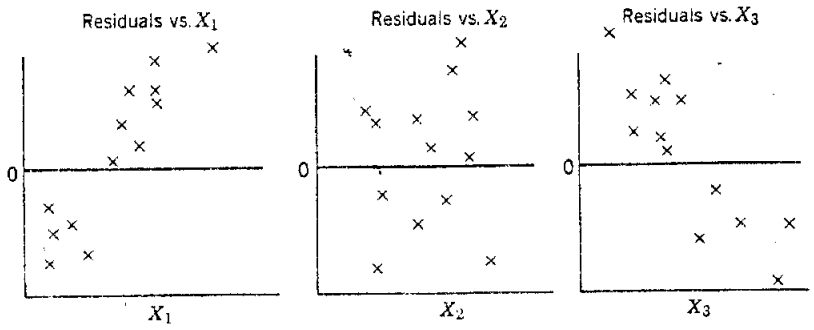
Residual Analysis for $\hat{X}_5 = f(X_4)$

Obs. No.	Observed Y	Predicted Y	Residual	Normal Deviate
1	78.5000300	73.2782200	5.2217800	.5825343
2	74.3000000	79.1835100	-4.8835100	-.5447974
3	104.3000000	102.6047000	1.4953000	.1668126
4	87.6000000	82.8743200	4.7258800	.5271901
5	95.9000000	93.2085900	2.6914100	.3002498
6	109.2000000	101.3283700	7.8716300	.8781478
7	102.7000000	113.1389600	-10.4389600	-1.1645554
8	72.5000000	85.0888100	-12.6888100	-1.4043896
9	93.1000000	101.3283700	-8.2283700	-.9179452
10	115.9000000	98.3757200	17.5242800	1.9549835
11	83.8000000	92.4704300	-8.6704300	-.9672608
12	113.3000000	108.7099900	4.5900100	.5120549
13	109.4000000	108.7099900	.6900100	.0769765

Residual Analysis for $\hat{X}_5 = f(X_3)$

Obs. NO.	Observed Y	Predicted Y	Residual	Normal Deviate
1		102.6679700	-24.1679700	-1.6201313
2	78.5000000 74.3000000	91.3659400	-17.0859400	-1.2852652
3		100.1564100	4.1435900	.3120609
5	104.3000000 87.6000000	100.1564100	-12.5564100	-.9456448
6	109.2000000 95.9000000	102.6679700	-6.7679700	-.5097075
		98.9006200	10.2993800	.7756640
7	102.7000000	88.6543700	13.6456300	1.0427381
8	72.5000000	82.5754700	-10.0754700	-.7588010
9	93.1000000	87.5985900	5.5014100	.4143206
10	115.9000000	105.1795300	10.7204700	.8073770
11	83.8000000	81.3196900	2.4803100	1867964
12	113.3000000	98.9006200	14.3993800	1: 0944420
13	109.4000000	100.1564100	9.2435900	.6961507

ขั้นตอนที่ 3 ใช้ค่าความคลาดเคลื่อน Z_i เป็นตัวแปรตามตัวใหม่ และเลือกตัวแปร X ที่ยังคงมีความสัมพันธ์สูงที่สุดกับ Z_i จากการ plot กราฟ ดังนี้



ขั้นตอนที่ 4 ถึงแม้ว่า X_1 และ X_3 มีความสัมพันธ์กับ Z_i พอ ๆ กัน เราคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ จะได้ X_1 มีค่ามากที่สุด ดังนั้นเราจะหาสมการถดถอยของ Z_i บน X_1 ซึ่งมีข้อมูลสำหรับการถดถอย คือ

Z_i	5.22	-4.88	1.50	4.73	2.69	7.87	-10.44	-12.59	-8.22	17.52	-8.67	4.59	0.69
X_1	7	1	11	11	7	11	3	1	2	21	1	11	10

ได้สมการถดถอย คือ $\hat{Z} = -10.10 + 1.35\ X_1$

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

$$Y_i = \hat{Y}_i + Z_i$$

$$Z_i = \hat{Z}_i + (Z_i - \hat{Z}_i)$$

หรืออาจเขียนได้ดังนี้ $Y_i = \hat{Y}_i + \hat{Z}_i + (Z_i - \hat{Z}_i)$

ดังนั้นในขั้นตอนที่ 2 สมการที่เราสามารถแทนโดย $\hat{Y}_i + \hat{Z}_i$ คือ

$$\mathbf{117.57 - 0.74\ X_4 + 10.10 + 1.35\ X_1 = 107.47 - 0.74\ X_4 + 1.35\ X_1}$$

ค่าความคลาดเคลื่อนที่เราคาดหวัง ขณะนี้ คือ $Z_i - \hat{Z}_i = Y_i - \hat{Y}_i - \hat{Z}_i$

ตัวแปรอื่น ๆ เราสามารถนำเข้ามาได้คล้ายการดำเนินการในแต่ละขั้นตอน โดยนำค่าความคลาด

เคลื่อนที่ได้ใหม่กับตัวแปรตัวใหม่ถัดมาหาสมการถดถอย จนกระทั่งได้สมการถดถอยที่มีตัวแปรที่เหลือไม่มีนัยสำคัญ กระบวนการจึงจะสิ้นสุด โดยไม่นำตัวแปรตัวสุดท้ายเข้ามาในสมการ จากตัวอย่างนี้สิ้นสุดที่ขั้นตอนที่สอง จะเห็นว่าสมการสุดท้ายที่ไม่ใช่วิธีกำลังสองน้อยที่สุด จะรวมตัวแปร X_1 และ X_4 เมื่อเปรียบเทียบกับวิธีกำลังสองน้อยที่สุด สมการของตัวแบบนี้ คือ

$$\hat{Y} = 130.10 - 0.61 X_4 + 1.44 X_1$$

แบบฝึกหัด

1. ข้อมูลเก็บรวบรวมในช่วงเวลาที่เท่า ๆ กัน มีค่าดังนี้

ค่าสังเกตที่	1	2	3	4	5	6	7	8	9	10	11
Y	1	5	4	7	10	8	9	13	14	13	18
X	-5	-4	-3	-2	-1	0	1	2	3	4	5

1.1 จงหาสมการถดถอยของ Y บน X

1.2 จงทดสอบสมมติฐาน $H_0: \beta_1 = 0$ ที่ระดับ $\alpha = 0.05$ และสร้างตารางวิเคราะห์ความแปรปรวน

1.3 จงหาช่วงความเชื่อมั่นของ $\beta_1 = 0$ เมื่อ $\alpha = 0.05$

1.4 จงหาช่วงความเชื่อมั่นของค่าเฉลี่ยของ Y เมื่อ $X = 3$

1.5 จงหาช่วงความเชื่อมั่นของค่าผลต่างของค่าเฉลี่ยของ Y เมื่อ $X_1 = 3$ และค่าเฉลี่ยของ Y เมื่อ $X_2 = -2$

1.6 มีเครื่องชี้นำบอกหรือไม่ว่า มีตัวแบบอื่นที่ดีกว่าตัวแบบที่มีอยู่

2. จากข้อมูลดังต่อไปนี้

Mix Moisture (Coded) X	4.7	5.0	5.2	5.2	5.9	4.7	5.9	5.2	5.3	5.9	5.6	5.0
Density (Coded) Y	3	3	4	5	10	2	9	3	7	6	6	4

$\sum X = 63.6$ $\sum Y = 62$ $\sum X^2 = 339.18$ $\sum Y^2 = 390$ $\bar{X} = 5.3$ $\bar{Y} = 5.17$ $\sum x^2 = 2.10$ $\sum y^2 = 69.67$ $\sum XY = 339.1$ $\sum xy = 10.5$

จงหา

2.1 จงหาสมการถดถอยของ Y บน X

2.2 จงหา 95 % ช่วงความเชื่อมั่นของ β_1

2.2 จงหาสมการถดถอยของ X บน Y

3. จากข้อมูลต่อไปนี้

X_1	1	4	9	11	3	8	5	10	2	7	6
X_2	8	2	-8	-10	6	-6	0	-12	4	-2	-4
Y	6	8	1	0	5	3	2	-4	10	-3	5

3.1 จงหาสมการถดถอยเชิงพหุ (Multiple Regression Model)

3.2 จงสร้างตารางวิเคราะห์ความแปรปรวน

3.3 ใช้ $\alpha = 0.05$ ทดสอบนัยสำคัญของ overall regression

3.4 คำนวณค่า R^2 พร้อมทั้งอธิบายค่าความแปรปรวนจากตัวแปร X ทั้งสองว่าสามารถอธิบายค่า Y ได้มากน้อยเพียงใด

3.5 จากปัญหานี้ มีค่า $(X'X)^{-1}$ ดังนี้

$$\begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix}$$

ใช้ผลจากตารางการวิเคราะห์ความแปรปรวนของเมตริกซ์นี้ คำนวณค่า

3.5.1 ความแปรปรวนของ b_1

3.5.2 ความแปรปรวนของ b_2

3.5.3 ความแปรปรวนของค่าพยากรณ์ของ Y ค่าเดียว เมื่อ $X_1 = 3$, $X_2 = 5$

3.6 ประโยชน์ที่จะได้รับจากการหาสมการถดถอยของ Y บน X_2 ตัวเดียว เป็นอย่างไร และเมื่อ X_1 เข้ามาอยู่ในสมการถดถอย แล้ว X_2 มีผลต่อ Y อย่างมีนัยสำคัญหรือไม่

3.7 จงสรุปผลที่ได้

4. จาก 8 ค่าสังเกตที่ได้จากเงื่อนไขสำคัญของการอิ่มตัว (X_1) และมีค่า

Y	66	43	36	23	22	14	12	7.6
X_1	38	41	34	35	31	34	29	32
X_2	47.5	21.3	36.5	18	29.5	14.2	21	10

ได้ $\sum X_1 = 274$ $\sum Y = 223.6$ $\sum X_2 = 198$ $\sum X_1^2 = 9488$ $\sum X_2^2 = 5979.08$ $\sum Y^2 = 8911.76$ $\sum X_1 Y = 8049.2$ $\sum X_2 Y = 6954.7$ $\sum X_1 X_2 = 6875.6$

4.1 จงหาสมการถดถอยของ Y บน X_1 และ X_2

4.2 overall regression มีนัยสำคัญหรือไม่ ที่ระดับนัยสำคัญ $\alpha = 0.05$

4.3 ตัวแปร X_1 และ X_2 สามารถอธิบายค่า Y ได้มากน้อยเพียงใด

5. จากข้อมูลต่อไปนี้

X_1	130	174	134	191	165	194	143	186	139	188	175	156	190	178	132	148
X_2	190	176	205	210	230	192	220	235	240	230	200	218	220	210	208	225
Y	35	81.7	42.5	98.3	52.7	82	34.5	95.4	56.7	84.4	94.3	44.3	83.3	91.4	43.5	51.7

5.1 จงประมาณค่า a, b_1 และ b_2 พร้อมทั้งหาสมการถดถอย

5.2 ทดสอบ overall regression ว่ามีนัยสำคัญหรือไม่

5.3 ตัวแปรตัวหนึ่งมีประโยชน์มากกว่าอีกตัวหนึ่งหรือไม่ ในการพยากรณ์ค่า Y

6. ใช้ค่าสังเกต 17 ค่า ที่กำหนดให้ จงหา

6.1 สมการถดถอย $\hat{Y} = a + b_1 X_1 + b_2 X_2$

6.2 ทดสอบว่าควรจะรวมตัวแปร X_1 และ X_2 ไว้ในสมการถดถอยหรือไม่

X_1	17	19	20	21	25	27	28	30
X_2	42	45	29	93	34	98	9	73
Y	90	71	76	63	80	75	99	73

7. ถ้าดำเนินการโดยวิธี stepwise regression และ backward elimination ของข้อมูลต่อไปนี้ จง สรุปผลการวิเคราะห์ของท่านโดยอาศัย print out ที่ให้มาว่าสมการถดถอยที่ดีที่สุดมีรูปแบบอย่างไรทั้งสองวิธีการ เมื่อ

X_1 = อุณหภูมิเฉลี่ยต่อเดือน หน่วยเป็น องศา F

X_2 = ปริมาณการผลิต หน่วยเป็น million pounds

X_3 = จำนวนวัน ในการผลิตแต่ละเดือน

X_4 = จำนวนคนงานที่ใช้ในการผลิตของแต่ละเดือน

X_5 = ตัวเลขสุ่มสองหลัก

X_6 = Y = ปริมาณน้ำที่ใช้ในแต่ละเดือน หน่วยเป็นแกลลอน

ข้อมูลเดิม

ลำดับ	Y	X_1	X_2	X_3	X_4	X_5
1	3067	58.8	7107	21	129	52
2	2828	65.2	6373	22	141	68
3	2891	70.9	6796	22	153	29
4	2994	77.4	9208	20	166	23
5	3082	79.3	14792	25	193	40
6	3898	81.0	14564	23	189	14
7	3502	71.9	11964	20	175	96
8	3060	63.9	13526	23	186	94
9	3211	54.5	12656	20	190	54
10	3286	39.5	14119	20	187	37
11	3542	44.5	16691	22	195	42
12	3125	43.6	14571	19	206	22
13	3022	56.0	13619	22	198	28
14	2922	64.7	14575	22	192	7
15	3950	73.0	14556	21	191	42
16	4488	78.9	18573	21	200	33
17	3295	79.4	15618	22	200	92

Correlation Matrix ที่ได้ดังนี้

ตัวแปร	1	2	3	4	5	Y
1	1.0	-0.024	0.438	- 0.082	0.108	0.286
2	- 0.024	1.0	0.106	0.918	- 0.111	0.631
3	0.438	0.106	1.0	0.032	0.038	-0.089
4	- 0.082	0.918	0.032	1.0	- 0.159	0.413
5	0.108	-0.111	0.038	- 0.159	1.0	-0.066
Y	0.286	0.631	- 0.089	0.143	-0.066	1.0

1. ฟังก์ชัน $Y = X_6 = f(X_2)$

Variable Entering 2

R - SQ 39.78

ส่วนเบี่ยงเบนมาตรฐานของ residuals 358.0

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	1270172	1270172	9.91
Residual	15	1922459	128163.9	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
2	0.079889	0.025	9.91

Constant term 2273.088

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

2. ฟังก์ชัน $Y = X_6 = f(X_4)$

Variable Entering 4

R - SQ 17.08

ส่วนเบี่ยงเบนมาตรฐานของ residuals 420.1

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	545213	545213	3.089
Residual	15	2647418	176494.5	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
4	8.39	4.78	3.09

Constant term 1777.723

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

3. ฟังก์ชัน $Y = X_6 = f(X_1, X_2)$

Variable Entering 1 , 2

R - SQ 48.85

ส่วนเบี่ยงเบนมาตรฐานของ residuals 341.5

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	1559525	779762.5	6.68
Residual	14	1633106	116650.4	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	9.96	6.32	2.48
2	0.08	0.02	11.13

Constant term 1615.495

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

4. ฟังก์ชัน $Y = X_6 = f(X_2, X_4)$

Variable Entering 2 , 4

R - SQ 57.42

ส่วนเบี่ยงเบนมาตรฐานของ residuals 311.604

ANOVA

sov.	d.f.	s s	M S	F
Regression	2	1833271	916635.5	9.44
Residual	14	1359360	97097. 14	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
2	0.203	0.056	13.27
4	- 21.567	8.956	5.80

Constant term 4600.806

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															
ลำดับ	16	17													
ค่าพยากรณ์															
e_t															

5. ฟังก์ชัน $Y = X_6 = f(X_1, X_2, X_3)$

Variable Entering 1 , 2 ,3

R - SQ 59.287

ส่วนเบี่ยงเบนมาตรฐานของ residuals 316.21

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	1892802	630934	6.3
Residual	13	1299929	99986.5	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	15.234	6.53	5 45
2	0.086	0.02	14.52
3	- 110.661	60.61	3.33

Constant term 3580.328

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

6. ฟังก์ชัน $Y = X_6 = f(X_1, X_2, X_4)$

Variable Entering 1 , 2 , 4

R - SQ 63.19

ส่วนเบี่ยงเบนมาตรฐานของ residuals 300.66

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	2017440	672480	7.44
Residual	13	1175191	90399.3	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	8.036	5.63	2.04
2	0.193	0.05	12.65
4	- 19.676	8.74	5.07

Constant term 3865.94

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

7. ฟังก์ชัน $Y = X_6 = f(X_1, X_2, X_3, X_4)$

Variable Entering 1, 2, 3, 4

R - SQ 76.70

ส่วนเบี่ยงเบนมาตรฐานของ residuals 248.96

ANOVA

SOV.	d.f.	S S	M S	F
Regression	4	2448834	612208.5	9.88
Residual	12	743797	61983.08	
Total	16	3192631		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	13.87	5.16	7.22
2	0.21	0.05	21.61
3	-126.69	48.02	6.96
4	- 21.82	7.29	8.97

Constant term 6360.339

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17
ค่าพยากรณ์		
e_t		

8. จากเซตของข้อมูลที่กำหนดให้ต่อไปนี้ จงวิเคราะห์หาสมการถดถอยที่ดีที่สุด โดยวิธี

1. all possible regression
2. stepwise regression

เมื่อ X_1 = อัตราการผลิต

X_2 = อุณหภูมิของน้ำที่ทำให้เย็นในขดของไนตริกออกไซด์

X_3 = การรวมตัวของ HNO_3 ในการดูดของเหลว

Y = เปอร์เซนต์ของ NH_3

ข้อมูลจากกระบวนการผลิตของการผสมออกซิเจนกับแอมมิโนเป็นไนตริกแอซิด

ข้อมูลเดิม

ลำดับ	Y	X_1	X_2	X_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

Correlation Matrix ที่ได้ดังนี้

ตัวแปร	Y	1	2	3
Y	1.0	0.9197	0.8755	0.3998
1	0.9197	1.0	0.7891	0.5001
2	0.8755	0.7819	1.0	0.3909
3	0.3998	0.5001	0.3909	1.0

1. ฟังก์ชัน $Y = X_4 = f(X_1)$

Variable Entering 1

R - SQ 84.578

ส่วนเบี่ยงเบนมาตรฐานของ residuals 4.098

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	1750.12	1750.12	104.199
Residual	19	319.12	16.796	
Total	20	2069.24		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	1.02	0.099	104.2

Constant term - 44.132

ตัวแปร	Square of Partials
2	0.408
3	0.031

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_i															
ลำดับ	16	17	18	19	20	21									
ค่าพยากรณ์															
e_i															

2. ฟังก์ชัน $Y = X_4 = f(X_2)$

Variable Entering 2

R - SQ 76.65

ส่วนเบี่ยงเบนมาตรฐานของ residuals 5.04

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	1586.087	1586.087	62.373
Residual	19	483.15	25.429	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
2	8.817	0.3567	62.37

Constant term - 41.91

ตัวแปร	Square of Partials
1	0.609
3	0.017

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17	18	19	20	21
ค่าพยากรณ์						
e_t						

3. ฟังก์ชัน $Y = X_4 = f(X_3)$

Variable Entering 3

R - SQ 15.986

ส่วนเบี่ยงเบนมาตรฐานของ residuals 9.565

ANOVA

SOV.	d.f.	S S	M S	F
Regression	1	330.795	330.795	3.615
Residual	19	1738.442	91.497	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
3	0.759	0.399	3.62

Constant term - 47.96

ตัวแปร	Square of Partials
1	0.822
2	0.726

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															
ลำดับ	16	17	18	19	20	21									
ค่าพยากรณ์															
e_t															

4. ฟังก์ชัน $Y = X_4 = f(X_1, X_2)$

Variable Entering 1 , 2

R - SQ 90.876

ส่วนเบี่ยงเบนมาตรฐานของ residuals 3.24

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	1880.442	940.221	89.6
Residual	18	188.795	10.49	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	0.671	0.1267	28.06
2	1.295	0.3674	12.42

Constant term - 50.359

ตัวแปร	Square of Partials
3	0.053

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															
ลำดับ	16	17	18	19	20	21									
ค่าพยากรณ์															
e_t															

5. ฟังก์ชัน $Y = X_4 = f(X_1, X_3)$

Variable Entering 1 , 3

R - SQ 85.06

ส่วนเบี่ยงเบนมาตรฐานของ residuals 4.14

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	1760.0996	880.05	51.26
Residual	18	309.1734	17.17	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	1.065	0.1167	83.22
3	-0.152	0.1997	0.58

Constant term - 33.686

ตัวแปร	Square of Partials
2	0.4215

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															

ลำดับ	16	17	18	19	20	21
ค่าพยากรณ์						
e_t						

6. ฟังก์ชัน $Y = X_4 = f(X_2, X_3)$

Variable Entering 2 , 3

R - SQ 77.04

ส่วนเบี่ยงเบนมาตรฐานของ residuals 5.137

ANOVA

SOV.	d.f.	S S	M S	F
Regression	2	1594.179	797.09	30.20
Residual	18	475.058	26.39	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
2	2.732	0.395	47.87
3	0.129	0.233	0.31

Constant term - 51.236

ตัวแปร	Square of Partials
1	0.6236

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_t															
ลำดับ	16	17	18	19	20	21									
ค่าพยากรณ์															
e_t															

7. ฟังก์ชัน $Y = X_4 = f(X_1, X_2, X_3)$

Variable Entering 1 , 2 , 3

R - SQ 91.358

ส่วนเบี่ยงเบนมาตรฐานของ residuals 3.243

ANOVA

SOV.	d.f.	S S	M S	F
Regression	3	1890.407	630.14	59.90
Residual	17	178.830	10.52	
Total	20	2069.237		

ตัวแปร	ส.ป.ส.ของตัวแปร	ค่าความคลาดเคลื่อน	Partial F - test
1	0.7156	0.13	28.16
2	1.2953	0.37	12 .39
3	-0.1521	0.16	0.95

Constant term - 39.92

สมการของการถดถอย คือ.....

ลำดับ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ค่าพยากรณ์															
e_i															

ลำดับ	16	17	18	19	20	21
ค่าพยากรณ์						
e_i						