บทที่ 9
การตรวจสอบและแก้ไขรูปแบบของตัวแบบเชิงเส้นพหุ

ในบทนี้จะกล่าวถึงการตรวจสอบรูปแบบที่เป็นปัญหาที่มักเกิดขึ้นกับการวิเคราะห์ตัวแบบเชิงเส้นพหุคือ ปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระ (multicollinearity) และปัญหาค่าผิดปกติทั้งตัวแปรอิสระและตัวแปรตาม รวมถึงปัญหาค่าที่มีอิทธิพลต่อตัวแบบ (influential) ปัญหาเหล่านี้อาจทำให้ผลการวิเคราะห์ข้อมูลผิดพลาดไปซึ่งส่งผลให้มีการตัดสินใจที่ผิดพลาดได้

9.1 ปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระ

ตัวแบบการถดถอยเชิงเส้นพหุเป็นตัวแบบที่มีตัวแปรอิสระจำนวนมากกว่า 1 ตัวดังนั้นอาจเป็นไปได้ที่ตัวแปรอิสระที่มีอยู่อาจมีความสัมพันธ์กันเองได้ซึ่งจะส่งผลในการเลือกตัวแปรอิสระที่มีความสำคัญต่อตัวแปรตามเนื่องจากจะพบว่าตัวแปรเหล่านั้นแสดงค่าความสัมพันธ์ต่อตัวแปรตาม หากนักวิจัยเก็บตัวแปรอิสระเหล่านั้นไว้ทั้งหมดในตัวแบบจะทำให้เกิดการซ้ำซ้อนและตัวแบบจะใหญ่เกินความจำเป็น ความสัมพันธ์ระหว่างตัวแปรอิสระเหล่านี้เรียก muliticollinearity เช่น หากต้องการพยากรณ์อัตราการใช้น้ำมันเชื้อเพลิงของรถยนต์ (Y) โดยใช้น้ำหนักบรรทุก (X_1) และกำลังม้าของเครื่องยนต์ (X_2) ซึ่งรถยนต์ที่สามารถบรรทุกได้มากจะมีกำลังม้าที่มากด้วยซึ่งทำให้มีอัตราการใช้น้ำมันที่สูง ดังนั้นตัวแปรอิสระทั้งสองมีความสัมพันธ์กันซึ่งไม่จำเป็นต้องใส่ตัวแปรทั้งสองในตัวแบบสามารถใช้เพียงตัวแปรใดตัวแปรหนึ่งก็สามารถพยากรณ์อัตราการใช้น้ำมันได้เป็นต้น วิธีหนึ่งในการตรวจสอบว่าตัวแปรแต่ละคู่มีความสัมพันธ์กันหรือไม่อย่างง่ายโดยการวาดแผนภาพกระจายระหว่างตัวแปรแต่ละคู่หรือการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient: r_{xy}) ของตัวแปรแต่ละคู่ หากค่า r_{xy} ไม่เท่ากับศูนย์แล้วแสดงว่าตัวแปรคู่นั้นมีความสัมพันธ์เชิงเส้นกัน

การที่ตัวแปรอิสระมีความสัมพันธ์ระหว่างกันนั้นจะส่งผลต่อการวิเคราะห์การถดถอยในหลายประเด็นดังนี้

1. ค่าผลรวมกำลังสองที่เพิ่มขึ้น (extra sum of squares) ของตัวแปรอิสระจะไม่ตรงกับความเป็นจริงเมื่อตัวแปรอิสระตัวนั้นมีความสัมพันธ์กับตัวแปรอิสระตัวอื่น หาก X_1 กับ X_2 ไม่มี

ความสัมพันธ์กันแล้วค่า $SSR(X_1)$ จะมีค่าเท่ากับ $SSR(X_1|\ X_2)$ นั่นคือ X_2 ไม่ส่งผลต่อค่าผลรวมกำลังสองถดถอยของ X_1 ทำให้ค่า b_1 ยังคงเท่าเดิม แต่หาก X_1 กับ X_2 มีความสัมพันธ์กันแล้วค่า $SSR(X_1)$ จะมีค่าไม่เท่ากับ $SSR(X_1|\ X_2)$ ส่งผลให้ค่า b_1 เปลี่ยนแปลงไป

2. ค่าพยากรณ์ที่ได้จะเปลี่ยนแปลงไปทั้งนี้เมื่อค่าประมาณของค่าสัมประสิทธิ์ของพารามิเตอร์ (b) ที่ได้เปลี่ยนแปลงไปส่งผลค่าสมการถดถอยเปลี่ยนแปลงไปจากความเป็นจริงทำให้การประมาณค่าสัมประสิทธิ์ของพารามิเตอร์และค่าพยากรณ์เปลี่ยนแปลงไปด้วย

3. ค่าความคลาดเคลื่อนมาตรฐานของค่าประมาณของค่าสัมประสิทธิ์ของพารามิเตอร์ ($se(b)$) จะมีค่ามากขึ้นจากความเป็นจริง

การตรวจสอบว่าตัวแปรทั้งสองมีความสัมพันธ์กันหรือไม่สามารถทำได้ดังนี้

1. ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรอิสระคู่หนึ่งคู่ใดมีความสัมพันธ์อย่างมีนัยสำคัญ

2. ค่าประมาณของค่าสัมประสิทธิ์ของพารามิเตอร์ (b) มีการเปลี่ยนแปลงค่าอย่างมากเมื่อมีการเพิ่มหรือลดตัวแปรอิสระหรือเมื่อมีการเปลี่ยนค่าหรือลบข้อมูลบางค่าออกไป

3. เมื่อการทดสอบตัวแปรอิสระที่จำเป็นต่อตัวแบบไม่พบว่ามีความสำคัญต่อตัวแบบ

4. ช่วงความเชื่อมั่นของค่าสัมประสิทธิ์ของพารามิเตอร์มีค่ากว้างมาก

5. ค่า *VIF* (variance inflation factor) ของตัวแปรอิสระหนึ่งมีค่ามากกว่า 10 นักวิจัยบางท่านใช้เกณฑ์ในการสรุปว่าตัวแปรมีความสัมพันธ์กันหากค่า *VIF* มีค่ามากกว่า 5

VIF เป็นการวัดค่าของความแปรปรวนของค่าประมาณของสัมประสิทธิ์ของพารามิเตอร์ที่เพิ่มขึ้นเมื่อตัวแปรอิสระมีความสัมพันธ์กัน โดยค่า *VIF* สามารถคำนวณได้ดังนี้

$$VIF_i = \frac{1}{1-R_i^2} \qquad i = 1,2,\ldots,k \tag{9.1}$$

โดย R_i^2 = ค่าสัมประสิทธิ์การตัดสินใจของตัวแบบที่ไม่รวมตัวแปรอิสระตัวที่ i

หากตัวแปรอิสระไม่มีความสัมพันธ์กันเชิงเส้นแล้วค่า R_i^2 จะมีค่าเป็น 0 ส่งผลให้ค่า *VIF* มีค่าเท่ากับ 1 แต่เมื่อตัวแปรมีความสัมพันธ์กันค่า R_i^2 เพิ่มขึ้นทำให้ค่า *VIF* สูงขึ้น

ข้อจำกัดของการใช้ *VIF* คือ *VIF* สามารถใช้หาตัวแปรอิสระที่มีความสัมพันธ์กันทีละ 2 ตัว หากตัวแปรอิสระมีความสัมพันธ์กันเป็นกลุ่มแล้วจะไม่สามารถใช้ *VIF* ในการตรวจสอบได้

6. ค่า *TOL* (tolerance) เป็นส่วนกลับของค่า *VIF* หรือ

$$TOL_i = \frac{1}{VIF_i} = 1 - R_i^2 \tag{9.2}$$

โปรแกรมสำเร็จรูปบางโปรแกรมจะแสดงค่า *TOL* โดยค่า *TOL* ที่นิยมใช้คือ 0.01 0.001 หรือ 0.0001 หากตัวแปรใดมีค่า *TOL* เกินเกณฑ์ดังกล่าวแสดงว่าตัวแปรมีความสัมพันธ์กัน

วิธีแก้ไขสามารถทำได้โดยการเลือกตัวแปรอิสระที่มีความสัมพันธ์กันเพียง 1 ตัวเข้าในตัวแบบเนื่องจากตัวแปรที่เหลือสามารถอธิบาย

หมายเหตุ

1. ในการคำนวณสมการถดถอยแบบเส้นโค้งนั้นจะมีการเพิ่มพจน์ที่มีกำลังสองหรือมีกำลังขั้นสูงของตัวแปรอิสระ เช่น $y = b_0 + b_1 x_1 + b_2 x_1^2$ เป็นต้น จะพบว่าพจน์ x_1^2 มีความสัมพันธ์เชิงเส้นกับ x_1 วิธีแก้ไขให้ลบค่า x_1 และ x_1^2 ทุกค่าด้วย $\bar{x}$ ดังนี้ $(x_1 - \bar{x})$ และ $(x_1 - \bar{x})^2$
2. หากใช้โปรแกรม MINITAB ช่วยในการคำนวณค่า *VIF* สามารถทำได้ดังนี้

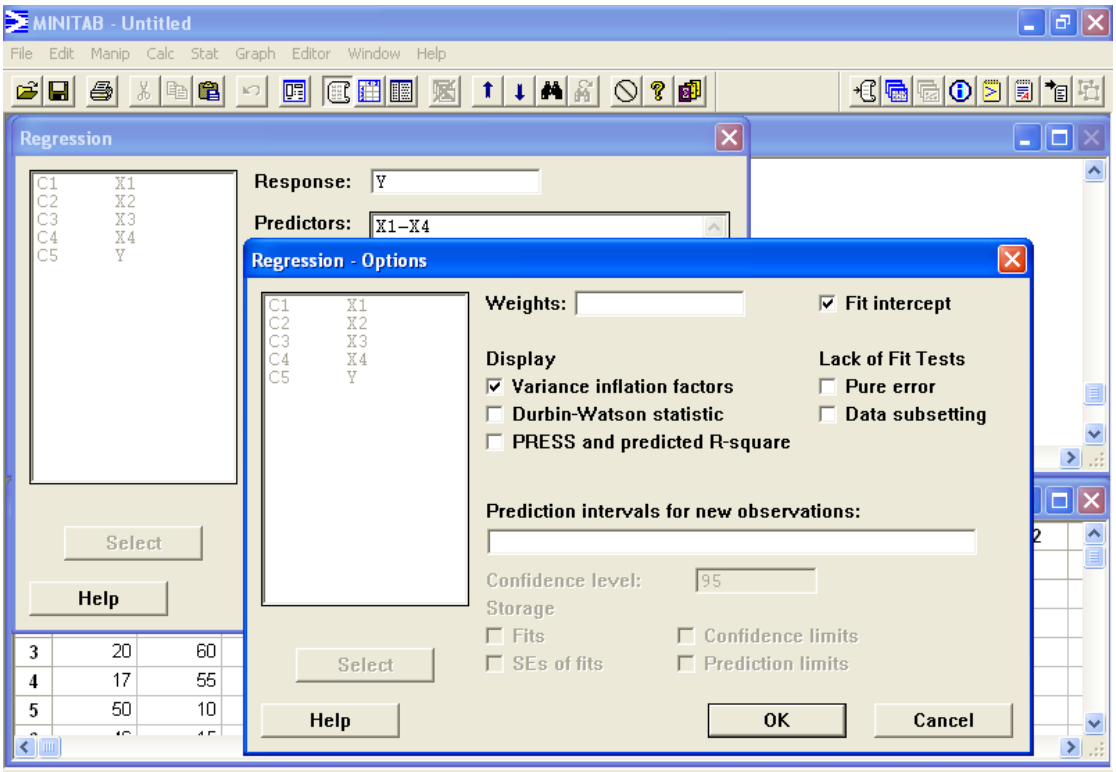
(1) เลือก “Stat” ที่เมนูบาร์

(2) เลือก “Regression”

(3) ระบุตัวแปรตามใน “Response:” และตัวแปรอิสระใน “Predictors:”

(4) เลือก “Options…” จากนั้นคลิกเลือก “Variance inflation factors” จากนั้นคลิก “OK”

(5) ผลลัพธ์ที่ได้จะแสดงในหน้าจอ “sessions” ดังภาพที่ 9.1



ภาพที่ 9.1 หน้าจอการคำนวณค่า *VIF*

3. ในงานวิจัยทางสังคมศาสตร์บางครั้งตัวแปรอิสระที่ได้มาจากการที่ผู้ให้ข้อมูลแสดงความคิดเห็นในแบบสอบถามในประเด็นต่างๆ ด้วยมาตรวัดแบบลิเคร์ท (Likert's scale) นั้นตัวแปรอิสระเหล่านั้นอาจมีความสัมพันธ์ซึ่งกันและกัน วิธีแก้ปัญหาอาจทำโดยการใช้การวิเคราะห์องค์ประกอบ (factor analysis) เพื่อรวมตัวแปรอิสระเหล่านั้นและใช้วิธีหมุนแกนแบบตั้งฉาก (orthogonal rotation) เพื่อให้ตัวแปรอิสระใหม่ที่ได้ไม่มีความสัมพันธ์กันและนำค่าคะแนนขององค์ประกอบ (factor score) ไปใช้ในการวิเคราะห์ (Mooi & Sarstedt, 2011, p. 169)

ตัวอย่างที่ 9.1 จากข้อมูลข้างล่างจงตรวจสอบว่าข้อมูลมีปัญหา multicollinearity หรือไม่

X_1	X_2	X_3	X_4	Y
11	470	16	2463	56
40	134	10	2048	67
20	60	11	3940	100
17	55	35	6505	155
50	10	36	5723	160
46	15	24	11520	161
60	18	45	5779	190
55	17	48	5969	210
95	28	75	8461	231
120	350	180	20106	350
94	280	60	13313	357
129	390	100	10771	374
125	370	125	15543	403
250	760	150	36194	1035
410	125	170	34703	1174
460	140	330	39204	1542

วิธีทำ

1. ตรวจสอบโดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์โดยใช้โปรแกรม MINITAB ได้ผลลัพธ์ดังนี้

Correlations: X1, X2, X3, X4, Y

	X1	X2	X3	X4
X2	0.205			
	0.447			
X3	0.897	0.278		
	0.000	0.296		
X4	0.934	0.400	0.886	
	0.000	0.124	0.000	
Y	0.982	0.288	0.898	0.961
	0.000	0.279	0.000	0.000

Cell Contents: Pearson correlation
P-Value

จากค่าสหสัมพันธ์ข้างต้นพบว่า X_1 มีความสัมพันธ์เชิงเส้นกับ X_3 และ X_4 ด้วยค่า r_{xy} เท่ากับ 0.897 และ 0.934 ตามลำดับ X_3 มีความสัมพันธ์เชิงเส้นกับ X_4 ด้วยค่า r_{xy} เท่ากับ 0.886 ดังนั้นข้อมูลชุดนี้มีปัญหา multicollinearity

2. จากการวิเคราะห์สมการถดถอยโดยใช้ MINITAB พบว่าค่า p-value ของตัวแปรอิสระ X_3 มีค่าเท่ากับ 0.993 นำไปสู่ข้อสรุปที่ว่าตัวแปร X_3 ไม่สำคัญต่อตัวแบบซึ่งขัดแย้งกับค่า r_{xy} ระหว่าง X_3 กับ Y พบว่ามีค่าเท่ากับ 0.898

นอกจากนี้หากพิจารณาค่า *VIF* จะพบว่าค่า *VIF* ของตัวแปรแต่ละตัวพบว่าค่า *VIF* ของ X_1 กับ X_4 มีค่าเท่ากับ 12.6 และ12.8 ตามลำดับซึ่งมากกว่า 10 แสดงว่าข้อมูลชุดนี้มีปัญหา multicollinearity

Regression Analysis: Y versus X1, X2, X3, X4

The regression equation is
Y = - 30.3 + 2.25 X1 + 0.049 X2 - 0.005 X3 + 0.0110 X4

Predictor	Coef	SE Coef	T	P	VIF
Constant	-30.35	30.12	-1.01	0.335	
X1	2.2476	0.4940	4.55	0.001	12.6
X2	0.0487	0.1111	0.44	0.670	1.6
X3	-0.0046	0.5228	-0.01	0.993	5.7
X4	0.011003	0.005473	2.01	0.070	12.8

S = 72.94 R-Sq = 98.0% R-Sq(adj) = 97.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2854006	713501	134.11	0.000
Residual Error	11	58524	5320		
Total	15	2912529			

Source	DF	Seq SS
X1	1	2808382
X2	1	23060
X3	1	1061
X4	1	21503

9.2 ปัญหาค่าผิดปกติในตัวแปรตาม

ค่าผิดปกติในข้อมูลนั้นมักมีผลต่อสมการถดถอยอาจทำให้ค่าประมาณของสัมประสิทธิ์ผิดไปจากความเป็นจริงและส่วนเหลือของค่านั้นมีค่าสูงผิดปกติ ดังนั้นการตรวจสอบและกำจัดค่าผิดปกติจึงจำเป็นที่ต้องตรวจสอบเพื่อให้ผลการวิเคราะห์ถูกต้อง ค่าผิดปกติสามารถพบได้ทั้งค่าผิดปกติในตัวแปรอิสระและในตัวแปรตาม

การวิเคราะห์หาค่าผิดปกติในตัวแปรเพียง 1 หรือ 2 ตัวนั้นสามารถวิเคราะห์ง่ายๆ โดยการวาดแผนภาพชนิดต่างๆ ของตัวแปรอิสระและตัวแปรตามรวมถึงแผนภาพของส่วนเหลือเช่นเดียวกับที่กล่าวมาแล้วในบทที่ 3 เพื่อช่วยในการตรวจสอบแต่เมื่อตัวแปรเพิ่มขึ้นการวาดแผนภาพจะยุ่งยาก การวิเคราะห์ค่าผิดปกติของตัวแปรอิสระจะใช้ส่วนเหลือในการวิเคราะห์ดังนี้

9.2.1 ส่วนเหลือมาตรฐาน

การวิเคราะห์โดยใช้ส่วนเหลือมาตรฐานมีวิธีเช่นเดียวกับการวิเคราะห์ข้อมูลที่มีตัวแปรอิสระเพียงตัวเดียวดังที่กล่าวมาแล้วในบทที่ 3 โดยมีสูตรดังนี้

$$e_i^* = \frac{e_i}{\sqrt{MSE}} \tag{9.3}$$

ข้อมูลที่ผิดปกติจะมีค่าส่วนเหลือมาตรฐานที่สูง นักสถิติบางท่านถือว่าหากข้อมูลมีค่า $|e_i^*|$ มากกว่า 2 แล้วถือว่าข้อมูลลำดับนั้นเป็นข้อมูลที่ผิดปกติ

9.2.2 ส่วนเหลือปรับแล้ว

ส่วนเหลือปรับแล้วในที่นี้จะมีสูตรการคำนวณที่แตกต่างจากในบทที่ 3 ดังนี้

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \tag{9.4}$$

โดย e_i = ส่วนเหลือของข้อมูลลำดับที่ i

MSE = ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน

h_{ii} = ค่า leverage ของข้อมูลลำดับที่ $i = \mathbf{X_i'(X'X)^{-1}X_i}$

เช่นเดียวกับส่วนเหลือมาตรฐานคือข้อมูลที่ผิดปกติจะมีค่าส่วนเหลือปรับแล้วสูง

หมายเหตุ

1. โปรแกรม MINITAB เรียกส่วนเหลือปรับแล้วว่า Standardized residual และถือว่าหากค่า r_i มีค่ามากกว่า 2 ถือว่าข้อมูลมีความผิดปกติ

2. ค่า r_i ไม่มีการแจกแจงแบบ t

9.2.3 ส่วนเหลือที่ถูกตัดออก

ส่วนเหลือที่ถูกตัดออก (deleted residual) สามารถหาได้โดยการตัดข้อมูลตำแหน่งที่ i หรือ y_i แล้วสร้างสมการถดถอยเพื่อพยากรณ์ข้อมูล y นั้นหรือ $\hat{y}_{i(i)}$ หากค่าพยากรณ์มีค่าใกล้เคียงกับค่าจริงแสดงว่าข้อมูลปกติแต่หากข้อมูลนั้นเป็นค่าผิดปกติแล้วค่าพยากรณ์จะแตกต่างจากค่าจริงมากเนื่องจากการตัดข้อมูลที่ผิดปกติออกไปทำให้สมการถดถอยที่ได้เปลี่ยนแปลงไป กำหนดให้ค่าความแตกต่างระหว่างค่าจริงและค่าพยากรณ์คือ d_i โดยมีสูตรคำนวณดังนี้

$$d_i = y_i - \hat{y}_{i(i)} \tag{9.5}$$

หรือเพื่อให้ง่ายแก่การคำนวณสามารถคำนวณโดย

$$d_i = \frac{e_i}{1-h_{ii}} \tag{9.6}$$

จะเห็นว่าเมื่อ h_{ii} มีค่าสูงแล้วจะทำให้ค่า d_i สูงขึ้นด้วย บ่อยครั้งจะพบว่าส่วนเหลือที่ถูกตัดออกจะช่วยในการหาค่าผิดปกติในตัวแปรอิสระที่ไม่สามารถตรวจพบในส่วนเหลือปกติ

นักวิจัยสามารถทดสอบว่าข้อมูลนั้นเป็นข้อมูลผิดปกติหรือไม่โดยใช้การทดสอบ t ตัวสถิติที่ใช้ในการทดสอบคือ

$$t = \frac{d_i}{\sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}} \tag{9.7}$$

โดย $MSE_{(i)}$ = ค่า MSE ที่ได้จากสมการถดถอยที่ไม่รวมข้อมูลลำดับที่ i เข้าในการคำนวณ

หากค่า $|t|$ มีค่ามากกว่า $t_{\alpha/2,n-p-1}$ แสดงว่าข้อมูลนั้นเป็นข้อมูลที่ผิดปกติ

9.2.4 ส่วนเหลือปรับแล้วที่ถูกตัดออก

ส่วนเหลือปรับแล้วที่ถูกตัดออก (Studentized deleted residual) มีวิธีการคำนวณที่แตกต่างจากสูตร (9.7) ดังนี้

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \tag{9.8}$$

หรือ

$$t_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{1/2} \tag{9.9}$$

ค่า t_i มีการแจกแจงแบบ t ที่มีองศาเสรีเท่ากับ n - p - 1 ดังนั้นข้อมูลที่มีค่า $|t_i|$ มากกว่า $t_{\frac{\alpha}{2},n-p-1}$ แสดงว่าข้อมูลนั้นเป็นข้อมูลที่ผิดปกติ นักวิจัยบางท่านใช้หลักการของ Bonferoni ในการทดสอบข้อมูลทั้ง n ค่าโดยใช้ค่าวิกฤตเป็น $t_{\frac{\alpha}{2n},n-p-1}$

หมายเหตุ

1. โปรแกรม MINITAB เรียกส่วนเหลือปรับแล้วที่ถูกตัดออกว่า deleted t residual

2. เมื่อ h_{ii} มีค่ามากขึ้นความแปรปรวนของส่วนเหลือจะมีค่าน้อยลงทั้งนี้เนื่องจากค่า h_{ii} ที่มากขึ้นทำให้ค่าพยากรณ์ $\hat{y}_i$ มีค่าใกล้เคียงกับค่า y_i มากขึ้นเมื่อข้อมูลมีค่า h_{ii} มากแล้วความแปรปรวนของส่วนเหลือจะมีค่าเป็น 0 นั่นคือค่าพยากรณ์มีค่าเท่ากับข้อมูลจริง

3. หากใช้โปรแกรม MINITAB ช่วยในการคำนวณสามารถทำได้ดังนี้

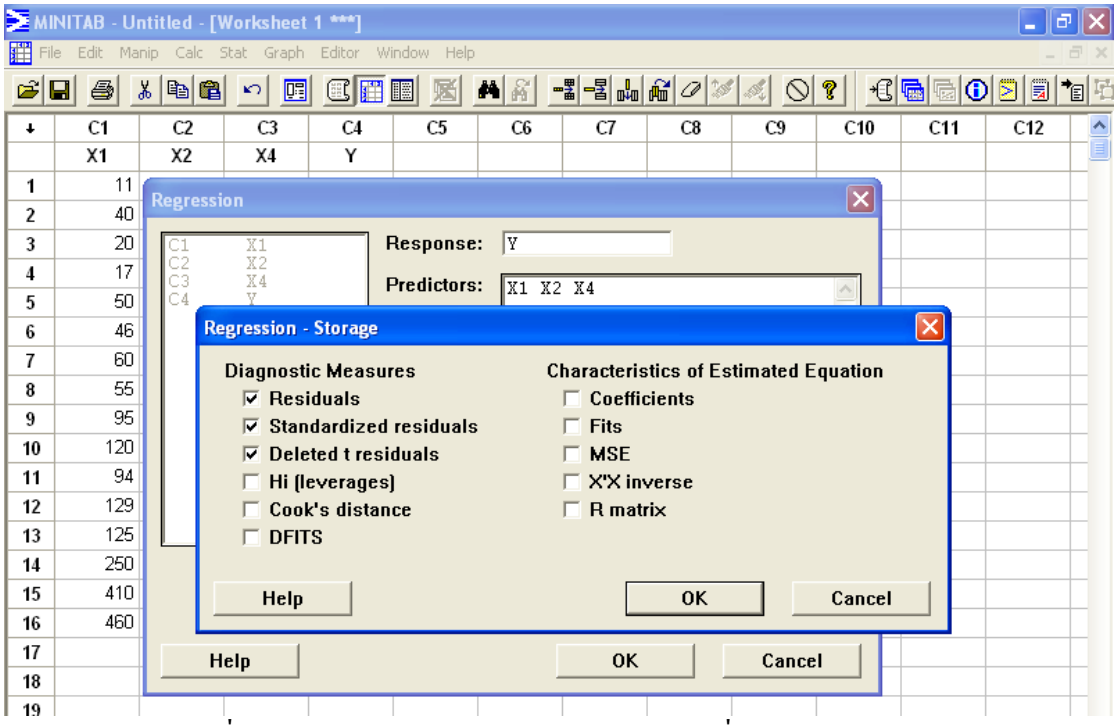
(1) เลือก “Stat” ที่เมนูบาร์

(2) เลือก “Regression”

(3) ระบุตัวแปรตามใน “Response:” และตัวแปรอิสระใน “Predictors:”

(4) เลือก “Storage…” จากนั้นคลิกเลือก “Residuals” หรือ “Standardized residuals” หรือ “Deleted t residuals” จากนั้นคลิก “OK”

(5) ผลลัพธ์ที่ได้จะเก็บในคอลัมน์สุดท้ายถัดจากข้อมูลใน “worksheet” ดังภาพที่ 9.2



ภาพที่ 9.2 หน้าจอการคำนวณส่วนเหลือปรับแล้วที่ถูกตัดออก

ตัวอย่างที่ 9.2 จากข้อมูล X_1 X_2 X_4 และ Y ที่ใช้ในตัวอย่างที่ 9.1 จงตรวจสอบว่ามีข้อมูลตัวแปรตามค่าใดที่ผิดปกติโดยใช้ส่วนเหลือมาตรฐาน ส่วนเหลือปรับแล้ว ส่วนเหลือที่ถูกตัดออกและส่วนเหลือปรับแล้วที่ถูกตัดออกโดยใช้โปรแกรม MINITAB

วิธีทำ

การวิเคราะห์การถดถอยโดยใช้ตัวแปรอิสระเพียง 3 ตัวโดยไม่รวมตัวแปร X_3 เนื่องจาก X_3 มีความสัมพันธ์กับ X_1 และ X_4 นอกจากนี้ยังพบว่า X_3 มีความสัมพันธ์กับตัวแปร Y น้อยกว่าตัวแปร X_1 และ X_4 ผลที่ได้จากโปรแกรม MINITAB พบว่าตัวแปร X_1 และ X_4 มีความสำคัญต่อตัวแปร Y นอกจากนี้ด้านล่างของผลลัพธ์ที่ได้ยังแสดงให้เห็นถึงข้อมูลที่มีค่าผิดปกติที่ได้จากการพิจารณาส่วนเหลือปรับแล้วหรือค่า “St Resid” ซึ่งพบว่าข้อมูลในลำดับที่ 10 และ 16 มีค่าส่วนเหลือปรับแล้วที่สูง

Regression Analysis: Y versus X1, X2, X4

The regression equation is
Y = - 30.4 + 2.25 X1 + 0.049 X2 + 0.0110 X4

Predictor	Coef	SE Coef	T	P
Constant	-30.39	28.52	-1.07	0.308
X1	2.2459	0.4356	5.16	0.000
X2	0.0487	0.1063	0.46	0.655
X4	0.010993	0.005111	2.15	0.053

S = 69.84 R-Sq = 98.0% R-Sq(adj) = 97.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	2854005	951335	195.06	0.000
Residual Error	12	58524	4877		
Total	15	2912529			

Source	DF	Seq SS
X1	1	2808382
X2	1	23060
X4	1	22563

Unusual Observations

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
10	120	350.0	477.2	32.3	-127.2	-2.05R
16	460	1542.0	1440.5	50.6	101.5	2.11R

R denotes an observation with a large standardized residual

เมื่อพิจารณาค่าส่วนเหลือ (e_i) ส่วนเหลือมาตรฐาน (e_i^*) ส่วนเหลือปรับแล้ว (r_i) และส่วนเหลือปรับแล้วที่ถูกตัดออก (t_i) พบว่าทั้งสี่วิธีให้ค่าที่สอดคล้องกันคือข้อมูลลำดับที่ 10 15 และ 16 มีค่าส่วนเหลือทั้ง 4 ชนิดที่สูงแสดงว่าข้อมูลทั้งสามมีแนวโน้มที่จะเป็นข้อมูลผิดปกติ ในที่นี้จะแสดงการคำนวณค่าต่างๆ โดยใช้ข้อมูลลำดับที่ 16

(1) ส่วนเหลือมาตรฐาน

$$e_{16}^* = \frac{e_{16}}{\sqrt{MSE}}$$

$$= \frac{101.501}{\sqrt{4877}} = 1.453429$$

(2) ส่วนเหลือปรับแล้ว

$$r_{16} = \frac{e_{16}}{\sqrt{MSE(1-h_{1616})}}$$

$$= \frac{101.501}{\sqrt{4877(1-0.524175)}} = 2.107028$$

(3) ส่วนเหลือที่ถูกตัดออก

$$d_{16} = \frac{e_{16}}{1-h_{1616}}$$

$$= \frac{101.501}{1-0.524175} \quad = \quad 213.3158$$

(4) ส่วนเหลือปรับแล้วที่ถูกตัดออก

$$t_{16} = \frac{e_{16}}{\sqrt{MSE_{(16}(1-h_{1616})}}$$

$$= \frac{101.501}{\sqrt{3352(1-0.524175}} \quad = \quad 2.541527$$

h_{ii}	e_i	e_i^*	r_i	d_i	t_i
0.417464	11.736	0.168052	0.22019	20.14639	0.21124
0.185583	-21.482	-0.30761	-0.34086	-26.3772	-0.32794
0.119177	39.239	0.561877	0.59869	44.54811	0.58196
0.173561	73.024	1.045657	1.15023	88.35982	1.16750
0.124177	14.695	0.210423	0.22485	16.7785	0.21573
0.332374	-39.288	-0.56258	-0.68853	-58.8473	-0.67263
0.118984	21.231	0.304014	0.3239	24.09831	0.31147
0.118769	50.421	0.721996	0.76911	57.21655	0.75522
0.110239	-46.344	-0.66362	-0.70353	-52.0859	-0.68792
0.213423	-127.171	-1.82101	-2.05324	-161.676	-2.44078
0.081383	16.302	0.233434	0.24356	17.74624	0.23377
0.299695	-22.715	-0.32526	-0.38868	-32.4359	-0.37449
0.106181	-36.214	-0.51856	-0.5485	-40.516	-0.53186
0.660903	69.059	0.988881	1.69818	203.6556	1.86540
0.413912	-103.995	-1.48914	-1.94516	-177.439	-2.25067
0.524175	101.501	1.453429	2.10703	213.3158	2.54152

เมื่อสร้างตัวแบบใหม่อีกครั้งโดยไม่รวมข้อมูลลำดับที่ 10 15 และ 16 ได้ตัวแบบใหม่ซึ่งมีสมการถดถอยที่แตกต่างจากเดิมดังนี้

Regression Analysis: Y versus X1, X2, X4

```
The regression equation is
Y = - 23.2 + 1.54 X1 + 0.0702 X2 + 0.0163 X4

Predictor        Coef     SE Coef          T        P
Constant       -23.19       18.30      -1.27    0.237
X1             1.5390      0.5058       3.04    0.014
X2            0.07019     0.07498       0.94    0.374
X4           0.016306    0.003677       4.43    0.002

S = 41.73       R-Sq = 98.0%     R-Sq(adj) = 97.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      770247      256749    147.41    0.000
Residual Error     9       15675        1742
Total             12      785922
```

9.3 ปัญหาค่าผิดปกติในตัวแปรอิสระ

ค่าผิดปกติในตัวแปรอิสระสามารถตรวจสอบโดยใช้ค่าที่อยู่ในแนวทแยงมุมหลักของ hat matrixหรือ h_{ii} โดยคุณสมบัติของค่า h_{ii} คือ

1. $0 \le h_{ii} \le 1$
2. $\sum_{i=1}^{n} h_{ii} = p$

h_{ii} เป็นตัววัดระยะทางระหว่างค่า X ทุกค่าในลำดับที่ i กับค่าเฉลี่ยของ X หาก h_{ii} มีค่าสูง แสดงว่า X ชุดนั้นอยู่ห่างค่าเฉลี่ยมาก ดังนั้นหากค่า h_{ii} มีค่ามากกว่าค่าเฉลี่ย ($\bar{h}$) เป็นสองเท่าแล้วจะถือว่าค่า h_{ii} มีขนาดใหญ่ โดยค่าเฉลี่ยสามารถคำนวณได้โดย

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n} \tag{9.10}$$

ดังนั้นหากค่า h_{ii} ใดมีค่ามากกว่า $\frac{2p}{n}$ แล้วข้อมูลชุดนั้นจัดเป็นค่าผิดปกติของตัวแปรอิสระ นักสถิติบางท่านกำหนดว่าเมื่อค่า h_{ii} มีค่ามากกว่า 0.5 แล้วถือว่าค่านั้นเป็นค่าที่ผิดปกติสูงแต่หากมีค่าอยู่ระหว่าง 0.2 ถึง 0.5 แล้วถือว่ามีความผิดปกติปานกลาง

ตัวอย่างที่ 9.3 จากข้อมูล X_1 X_2 X_4 และ Y ที่ใช้ในตัวอย่างที่ 9.1 จงตรวจสอบว่ามีข้อมูลตัวแปรอิสระค่าใดที่ผิดปกติโดยใช้ค่า h_{ii} ที่ได้จากโปรแกรม MINITAB

วิธีทำ

จากตารางในตัวอย่างที่ 3.2 พบว่าค่า h_{ii} ที่ได้จากข้อมูลลำดับที่ 1 14 15 และ 16 มีค่ามากกว่า 0.375 (หรือ $\frac{2p}{n}=\frac{2\times 3}{16}$) แต่เฉพาะข้อมูลลำดับที่ 14 และ 16 ที่มีค่ามากกว่า 0.5 แสดงว่าข้อมูลเหล่านี้มีความผิดปกติสำหรับตัวแปรอิสระ

9.4 ปัญหาค่าที่มีอิทธิพล

บางครั้งข้อมูลบางกลุ่มมีอิทธิพลต่อสมการถดถอยหรือค่าประมาณของค่าสัมประสิทธิ์หรือค่าพยากรณ์ให้มีค่าที่ผิดไปจากความเป็นจริง ส่วนใหญ่ค่าผิดปกติของตัวแปรอิสระและตัวแปรตามเป็นค่าที่มีอิทธิพลแต่ไม่จำเป็นเสมอไป หลักการของการตรวจหาข้อมูลที่มีอิทธิพลคือการละข้อมูลค่านั้นหรือกลุ่มนั้นออกจากการสร้างตัวแบบแล้วเปรียบเทียบตัวแบบที่ได้

9.4.1 ค่าที่มีอิทธิพลต่อค่าพยากรณ์หนึ่งค่า

ค่า $(DFFITS)_i$ เป็นค่าที่ใช้วัดอิทธิพลของข้อมูลลำดับที่ i ที่มีผลต่อค่าพยากรณ์ $(\hat{Y}_i)$ โดยการหาความแตกต่างระหว่างค่าพยากรณ์ที่มีข้อมูลลำดับนั้นกับค่าพยากรณ์ที่ไม่รวมข้อมูลลำดับนั้น

$$(DFFITS)_i=\frac{\hat{Y}_i-\hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} \tag{9.11}$$

$$=e_i\sqrt{\left(\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)}$$

$$=t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}} \tag{9.12}$$

โดย $\hat{Y}_i$ = ค่าพยากรณ์ในลำดับที่ i ที่ใช้ข้อมูลทั้งหมด

$\hat{Y}_{i(i)}$ = ค่าพยากรณ์ในลำดับที่ i ที่ใช้ข้อมูลทั้งหมดยกเว้นข้อมูลลำดับที่ i

t_i = ส่วนเหลือปรับแล้วที่ถูกตัดออก

จากสูตร (9.11) จะเห็นว่า $(DFFITS)_i$ ขึ้นอยู่กับค่า e_i และ h_{ii} หากข้อมูลมีความผิดปกติในตัวแปรอิสระแล้วค่า h_{ii} จะมีค่าสูงส่งผลให้ค่า $(DFFITS)_i$ สูงด้วย ในข้อมูลขนาดกลางและเล็กนั้นข้อมูลที่มีค่า $|(DFFITS)_i|$ มากกว่า 1 จัดเป็นค่าที่มีอิทธิพลแต่ในข้อมูลขนาดใหญ่นั้นหากมีค่า $|(DFFITS)_i|$ มากกว่า $2\sqrt{\frac{p}{n}}$ จึงจะจัดเป็นค่าที่มีอิทธิพล

9.4.2 ค่าที่มีอิทธิพลต่อค่าพยากรณ์ทั้งหมด

ค่า Cook's Distance หรือ D_i ใช้ในการตรวจสอบค่าที่มีอิทธิพลต่อค่าพยากรณ์ทั้งหมด

$$D_i = \frac{(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})'(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})}{pMSE} \tag{9.13}$$

$$= \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{9.14}$$

โดย $\hat{\boldsymbol{Y}}$ = เวคเตอร์ของค่าพยากรณ์ที่ใช้ข้อมูลทั้งหมดในการคำนวณ

$\hat{\boldsymbol{Y}}_{(i)}$ = เวคเตอร์ของค่าพยากรณ์ที่ไม่ใช้ข้อมูลลำดับที่ i ในการคำนวณ

ค่า D_i มีการแจกแจงแบบ F ที่มีองศาเสรีเท่ากับ p และ $n - p$ ค่า D_i ก็เช่นเดียวกับค่า $(DFFITS)_i$ ที่ขึ้นกับค่า e_i และ h_{ii} กล่าวคือ หากค่า e_i และ h_{ii} มีค่าสูงแล้วค่า D_i จะสูงตามด้วย หากค่าเปอร์เซ็นไทล์ของ ค่า D_i นั้นมีค่าน้อยกว่า 10 หรือ 20 แล้วข้อมูลนั้นจะมีอิทธิพลต่อค่าพยากรณ์ทั้งหมดเพียงเล็กน้อย หากค่าเปอร์เซ็นไทล์ของค่า D_i นั้นมีค่าตั้งแต่ 50 ขึ้นไปแล้วข้อมูลนั้นจะมีอิทธิพลต่อค่าพยากรณ์มากกล่าวคือ ค่าพยากรณ์ที่ได้จากการใช้ข้อมูลค่านั้นในการสร้างสมการถดถอยจะแตกต่างอย่างมากจากค่าพยากรณ์ที่ไม่ใช้ข้อมูลนั้น เช่น หากมีตัวแปรอิสระจำนวน 4 ตัว ข้อมูลจำนวน 25 ข้อมูลและมีค่า D_i เท่ากับ 1.25 จะมีค่าเท่ากับเปอร์เซ็นไทล์ของ $F_{(4,21)}$ ที่ 67.94 ซึ่งมีค่ามากกว่าเปอร์เซ็นไทล์ที่ 50 ดังนั้นข้อมูลค่านี้มีอิทธิพลต่อค่าพยากรณ์ทั้งหมด เป็นต้น

9.4.3 ค่าที่มีอิทธิพลต่อค่าสัมประสิทธิ์

ค่าที่มีอิทธิพลต่อค่าสัมประสิทธิ์เป็นค่าที่อาจทำให้ค่าประมาณของค่าสัมประสิทธิ์สูงขึ้นหรือต่ำลง โดยใช้ $(DFBETAS)_{k(i)}$ ในการหาค่าที่มีอิทธิพลต่อค่าสัมประสิทธิ์ทำโดยหาความแตกต่างระหว่างค่าประมาณของค่าสัมประสิทธิ์ที่มีข้อมูลลำดับนั้นกับค่าประมาณของค่าสัมประสิทธิ์ที่ไม่รวมข้อมูลลำดับนั้น

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}} \qquad ; \quad k = 0, 1, \ldots, p-1 \tag{9.15}$$

โดย b_k = ค่าประมาณของค่าสัมประสิทธิ์ของตัวแปรอิสระ x_k ที่ใช้ข้อมูลทั้งหมดในการคำนวณ

$b_{k(i)}$ = ค่าประมาณของค่าสัมประสิทธิ์ของตัวแปรอิสระ x_k ที่ไม่ใช้ข้อมูลลำดับที่ i ในการคำนวณ

c_{kk} = ค่าที่อยู่ในแนวทแยงมุมหลักลำดับที่ k ของ $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

หากข้อมูลมีอิทธิพลต่อค่าสัมประสิทธิ์ลำดับที่ k ของสมการถดถอยหรือสัมประสิทธิ์ของตัวแปรอิสระ x_k แล้วค่า $|(DFBETAS)_{k(i)}|$ จะสูงด้วย ในข้อมูลขนาดกลางและเล็กนั้นข้อมูลที่มีค่า $|(DFBETAS)_{k(i)}|$ มากกว่า 1 จัดเป็นค่าที่มีอิทธิพลแต่ในข้อมูลขนาดใหญ่นั้นหากมีค่า $|(DFBETAS)_{k(i)}|$ มากกว่า $2/\sqrt{n}$ จึงจะจัดเป็นค่าที่มีอิทธิพล

หมายเหตุ

1. หากใช้โปรแกรม MINITAB ช่วยในการคำนวณ $(DFFITS)_i$, Cook's Distance สามารถทำได้ดังนี้

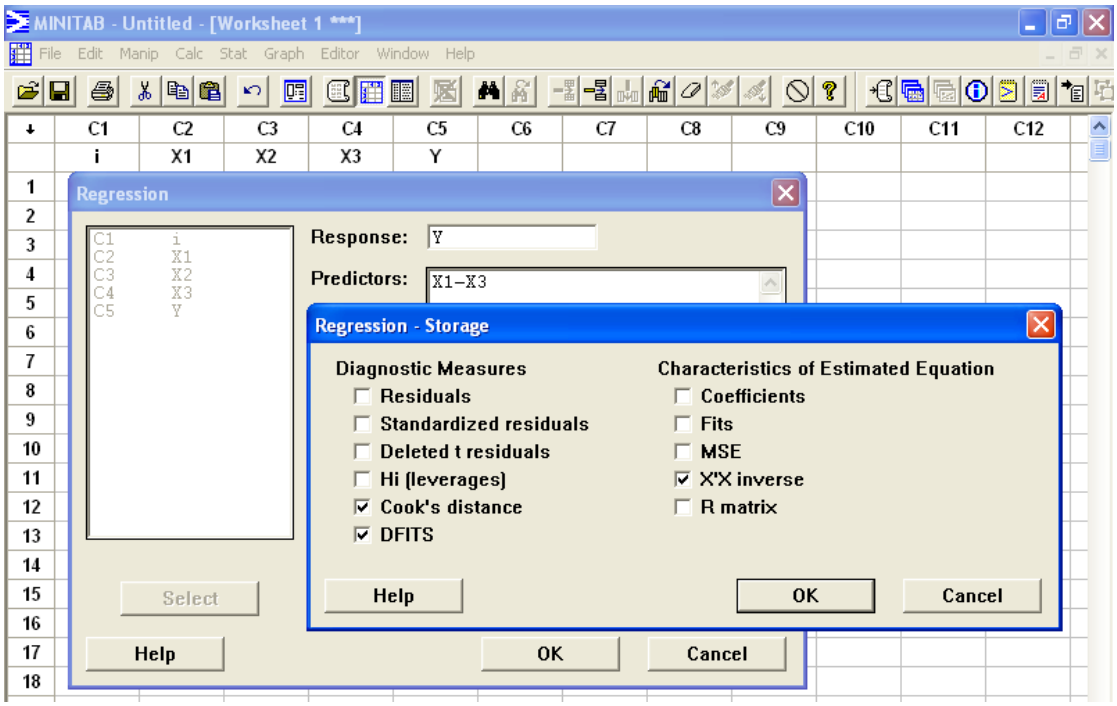
(1) เลือก "Stat" ที่เมนูบาร์

(2) เลือก "Regression"

(3) ระบุตัวแปรตามใน "Response:" และตัวแปรอิสระใน "Predictors:"

(4) เลือก "Storage…" จากนั้นคลิกเลือก "DFITS" และ "Cook's distance"จากนั้นคลิก "OK"

(5) ผลลัพธ์ที่ได้จะเก็บในคอลัมน์ที่ชื่อ "DFITS" และ "COOK" ใน "worksheet" ดังภาพที่ 9.3



ภาพที่ 9.3 หน้าจอการคำนวณ $(DFFITS)_i$ และ Cook's Distance

2. โปรแกรม MINITAB ไม่สามารถคำนวณค่า $(DFBETAS)_{k(i)}$ ได้แต่สามารถคำนวณค่า $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ได้โดยเลือกจาก “X’X inverse” ใน “Storage”

ตัวอย่างที่ 9.4 จากข้อมูลข้างล่างจงตรวจสอบว่ามีข้อมูลใดที่เป็นค่าที่มีอิทธิพลโดยใช้ค่า $(DFFITS)_i$ Cook’s Distance และ $(DFBETAS)_{k(i)}$ ที่ได้จากโปรแกรม MINITAB

i	X_1	X_2	X_3	Y
1	12.4	105.0	23.0	3.1
2	12.7	108.0	24.0	3.5
3	12.9	102.0	23.1	2.6
4	13.9	105.0	23.0	3.0
5	13.0	104.0	20.7	2.9
6	11.7	102.0	22.1	2.7
7	11.3	104.0	23.2	3.1
8	13.9	111.0	26.9	3.4
9	13.3	101.0	22.2	2.5
10	12.1	105.0	24.6	3.0

วิธีทำ

จากตัวแบบที่สร้างโดยใช้โปรแกรม MINITAB พบว่าตัวแปรอิสระทั้งสามไม่มีความสัมพันธ์กันเนื่องจากค่า VIF ทั้งหมดมีค่าน้อยกว่า 10

Regression Analysis: y versus X1, X2, X3

The regression equation is
y = - 8.09 - 0.0859 X1 + 0.127 X2 - 0.0487 X3

Predictor	Coef	SE Coef	T	P	VIF
Constant	-8.086	1.121	-7.22	0.000	
X1	-0.08586	0.03357	-2.56	0.043	1.2
X2	0.12700	0.01527	8.32	0.000	2.8
X3	-0.04873	0.02642	-1.84	0.115	2.6

S = 0.08138 R-Sq = 95.3% R-Sq(adj) = 93.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.81210	0.27070	40.87	0.000
Residual Error	6	0.03974	0.00662		
Total	9	0.85184			

Source	DF	Seq SS
X1	1	0.01580
X2	1	0.77376
X3	1	0.02254

เมื่อพิจารณาค่าที่มีอิทธิพลโดยวิธีทั้ง 3 ได้ผลดังนี้

(1) $(DFFIT)_i$

เนื่องจากข้อมูลมีขนาดเล็กดังนั้นค่า $|(DFFIT)_i|$ ที่มีค่ามากกว่า 1 จะเป็นข้อมูลที่มีอิทธิพลต่อค่าพยากรณ์ของข้อมูลลำดับนั้นพบว่าข้อมูลลำดับที่ 2 5 และ 8 ที่มีค่า $|(DFFIT)_i|$ ที่มีค่ามากกว่า 1 แสดงว่าข้อมูลทั้งสามค่าทีอิทธิพลต่อค่าพยากรณ์นั้นๆ ในที่นี้จะแสดงการคำนวณโดยใช้ข้อมูลลำดับที่ 8

$$(DFFIT)_8 = e_8\sqrt{\left(\frac{n-p-1}{SSE(1-h_{88})-e_8^2}\right)\left(\frac{h_{88}}{1-h_{88}}\right)}$$

$$= -0.9878\sqrt{\left(\frac{10-4-1}{0.0397\times(1-0.7304)-(-0.9878^2)}\right)\left(\frac{0.7304}{1-0.7304}\right)}$$

$$= -5.5711$$

(2) Cook's Distance

เนื่องจากข้อมูลมีขนาดเล็กดังนั้นค่า D_i ที่มีเปอร์เซ็นไทล์มากกว่า 10 แล้วข้อมูลนั้นเป็นค่าที่มีอิทธิพลโดยเทียบกับ $F_{(4,\ 6)}$ พบว่าข้อมูลลำดับที่ 2 5 และ 8 ที่มีเปอร์เซ็นไทล์เกิน 10 แสดงว่าข้อมูลทั้งสามค่าที่มีอิทธิพลต่อค่าพยากรณ์ทั้งหมด ในที่นี้จะแสดงการคำนวณโดยใช้ข้อมูลลำดับที่ 8

$$D_8 = \frac{e_8^2}{pMSE}\left[\frac{h_{88}}{(1-h_{88})^2}\right]$$
$$= \frac{(-0.0988^2)}{4\times 0.0066}\left[\frac{0.7304}{(1-0.7304)^2}\right]$$
$$= 2.8291$$

ค่า D_i = 2.8291 นั้นเท่ากับเปอร์เซ็นไทล์ของ $F_{(4,6)}$ ที่ 87.6845

(3) $(DFBETAS)_{k(i)}$

เนื่องจากข้อมูลมีขนาดเล็กดังนั้นค่า $|(DFBETAS)_{k(i)}|$ ที่มีค่ามากกว่า 1 จะเป็นข้อมูลที่มีอิทธิพลต่อค่าประมาณของค่าสัมประสิทธิ์พบว่าข้อมูลลำดับที่ 5 และ 8 ที่มีค่า $|(DFBETAS)_{k(i)}|$ ที่มีค่ามากกว่า 1 แสดงว่าข้อมูลทั้งสองค่าที่มีอิทธิพลต่อค่าพยากรณ์นั้นๆ ในที่นี้จะแสดงการคำนวณค่าของ b_0 โดยใช้ข้อมูลลำดับที่ 8

$$\mathbf{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 189.593 & -0.100 & -2.319 & 2.343 \\ -0.100 & 0.170 & -0.022 & 0.009 \\ -2.319 & -0.022 & 0.035 & -0.047 \\ 2.343 & 0.009 & -0.047 & 0.105 \end{bmatrix}$$

ดังนั้น

$$(DFBETAS)_{0(8)} = \frac{b_0 - b_{0(8)}}{\sqrt{MSE_{(8)}c_{00}}}$$
$$= \frac{-8.0860-(-10.3396)}{\sqrt{0.0020\times 189.593}}$$
$$= 3.6782$$

$(DFFIT)_i$	D_i	เปอร์เซ็นไทล์ของ D_i	$(DFBETAS)_{k(i)}$			
			b_0	b_1	b_2	b_3
0.0832	0.0021	0.0011	0.2141	0.5314	-0.3431	0.1400
1.3771	0.3137	14.0765	-0.6033	0.3482	0.4643	-0.3088
0.0554	0.0009	0.0002	0.0926	0.5117	-0.2271	0.0718
0.6992	0.1228	3.0994	-0.1870	1.2671	-0.1258	-0.1848
-2.7133	1.2507	61.6645	1.0060	0.2854	-1.7108	2.3717
-0.8365	0.1551	4.6348	-0.4035	1.0454	-0.0772	0.1920

$(DFFIT)_i$	D_i	เปอร์เซ็นไทล์	$(DFBETAS)_{k(i)}$			
		ของ D_i	b_0	b_1	b_2	b_3
0.4990	0.0698	1.1168	0.1779	-0.1272	-0.1529	0.1793
-5.5711	2.8291	87.6845	3.6782	-1.3588	-1.4796	-2.0120
0.3381	0.0334	0.2769	0.2186	0.2025	-0.2435	0.0992
-0.0992	0.0029	0.0023	-0.0335	0.0434	0.0361	-0.0656

เมื่อสร้างตัวแบบใหม่อีกครั้งโดยไม่รวมข้อมูลลำดับที่ 5 และ 8 ได้ตัวแบบใหม่ซึ่งมีสมการถดถอยที่แตกต่างจากเดิมดังนี้

Regression Analysis: Y versus X1, X2, X3

```
The regression equation is
Y = - 10.2 - 0.0562 X1 + 0.144 X2 - 0.0516 X3

Predictor        Coef     SE Coef          T        P
Constant     -10.1892      0.8345     -12.21    0.000
X1           -0.05615     0.02068      -2.71    0.053
X2            0.14448     0.01114      12.97    0.000
X3           -0.05164     0.03048      -1.69    0.166

S = 0.04563     R-Sq = 98.7%     R-Sq(adj) = 97.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     0.63202     0.21067    101.18    0.000
Residual Error     4     0.00833     0.00208
Total              7     0.64035
```

สรุป

การวิเคราะห์สมการถดถอยนั้นหากข้อมูลมีปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระ ปัญหาค่าผิดปกติทั้งตัวแปรอิสระและตัวแปรตาม รวมถึงปัญหาค่าที่มีอิทธิพลต่อตัวแบบแล้วอาจทำให้นักวิจัยสรุปผลผิดพลาดได้ ดังนั้นการตรวจสอบและแก้ไขปัญหาเหล่านี้จึงเป็นสิ่งที่สำคัญอย่างยิ่งเพื่อป้องกันการสรุปผลที่ผิดพลาด การแก้ไขปัญหาเหล่านี้มีด้วยกันหลายวิธีเช่น การตัดข้อมูลนั้นทิ้งไปหรือการแปลงข้อมูล เป็นต้น

คำถามท้ายบท

9.1 จงอธิบายว่าทำไมการที่ตัวแปรอิสระมีความสัมพันธ์กันจึงเป็นปัญหาต่อการวิเคราะห์สมการถดถอย

9.2 หากท่านทราบว่าตัวแปรอิสระมีความสัมพันธ์กันจะมีวิธีแก้ไขอย่างไร

9.3 ท่านจะสามารถทราบได้อย่างไรว่าตัวแปรอิสระมีความสัมพันธ์กันหรือไม่

9.4 จากค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระทั้ง 4 ตัวที่ได้จากโปรแกรม MINITAB จงอธิบายว่าตัวแปรทั้ง 4 มีปัญหา multicollinearlity หรือไม่

Correlations: x1, x2, x3, x4

	x1	x2	x3
x2	0.823		
	0.000		
x3	0.494	0.468	
	0.012	0.018	
x4	0.957	0.774	0.470
	0.000	0.000	0.018

Cell Contents: Pearson correlation
P-Value

9.5 จากข้อมูลข้างล่างจงสร้างตัวแบบกำลังสองหรือ $y = b_0 + b_1 x_1 + b_2 x_1^2$ โดยต้องหลีกเลี่ยงปัญหา multicollinearlity

X	420	380	480	340	465	460	430	370	390
Y	24	21	31	21	29	26	25	21	22

9.6 จงอธิบายความแตกต่างระหว่างส่วนเหลือชนิดต่างๆ

9.7 ค่า h_{ii} ใช้บอกค่าผิดปกติในตัวแปรอิสระและตัวแปรตามได้อย่างไร

9.8 จากข้อ 8.9 จงหาว่าข้อมูลชุดนี้มีค่าผิดปกติในตัวแปรตามหรือไม่โดยใช้ส่วนเหลือ ส่วนเหลือมาตรฐาน ส่วนเหลือปรับแล้วและส่วนเหลือปรับแล้วที่ถูกตัดออกพร้อมทั้งอธิบายว่าผลการวิเคราะห์ที่ได้แตกต่างกันหรือไม่ อย่างไร หากพบว่าค่าผิดปกติจงตัดข้อมูลนั้นทิ้งแล้วสร้างสมการถดถอยใหม่พร้อมทั้งอธิบายว่าสมการถดถอยที่ได้แตกต่างจากเดิมหรือไม่

9.9 จากข้อ 8.9 จงตรวจสอบว่ามีข้อมูลตัวแปรอิสระค่าใดที่ผิดปกติโดยใช้ค่า h_{ii}

9.10 จากข้อ 8.9 จงตรวจสอบว่ามีข้อมูลใดที่เป็นค่าที่มีอิทธิพลโดยใช้ค่า $(DFFITS)_i$ Cook's Distance และ $(DFBETAS)_{k(i)}$

9.11 ในการหาความสัมพันธ์ระหว่างความแข็งแรงและความยืดหยุ่นของไม้ที่มีต่อความหนาแน่นของไม้ จงวิเคราะห์ว่าข้อมูลชุดนี้มีค่าผิดปกติในตัวแปรตามหรือไม่โดยใช้ส่วนเหลือแบบต่างๆ ช่วยในการวิเคราะห์

ความแข็งแรง	ความยืดหยุ่น	ความหนาแน่น
1000	99	25.3
1112	173	28.2
1033	188	28.6
1087	133	29.1
1069	146	30.7
925	91	31.4
1306	188	32.5
1306	194	36.8
1323	195	37.1
1379	177	38.3
1332	182	39.0
1254	110	39.6
1587	203	40.1
1145	193	40.3
1438	167	40.3
1281	188	40.6
1595	238	42.3
1129	130	42.4
1492	189	42.5
1605	213	43.0
1647	165	43.0
1539	210	46.7
1706	224	49.0
1728	228	50.2
1703	209	50.3

ความแข็งแรง	ความยืดหยุ่น	ความหนาแน่น
1897	240	50.3
1822	248	51.3
2129	261	51.7
2053	245	52.8
1676	186	53.8
1621	188	53.9
1990	252	54.9
1764	222	55.1
1909	244	55.2
2086	274	55.3
1916	276	56.9
1889	254	57.3
1870	238	58.3
2036	264	58.6
2570	189	58.7
1474	223	59.5
2116	245	60.8
2054	272	61.3
1994	264	61.5
1746	196	63.2
2604	268	63.3
1767	205	68.1
2649	346	68.9
2159	246	68.9
2078	235	70.8

ที่มา: Brown.& Maritz,1982, p. 318-331

9.12 จากข้อ 9.11 มีข้อมูลใดที่เป็นค่าที่มีอิทธิพลพร้อมทั้งอธิบายผลที่ได้ หากพบว่ามีข้อมูลที่มีค่าที่มีอิทธิพลจงตัดข้อมูลนั้นทิ้งแล้วสร้างสมการถดถอยใหม่พร้อมทั้งอธิบายว่าสมการถดถอยที่ได้แตกต่างจากเดิมหรือไม่

9.13 จากข้อมูลข้างล่างจงตรวจสอบว่าตัวแปรอิสระมีความสัมพันธ์กันหรือไม่และท่านจะแก้ปัญหานี้อย่างไรในการสร้างสมการถดถอย

X_1	X_2	X_3	X_4	X_5	Y
16	25	473	18	44.5	566
44	20	1340	10	69.2	696
21	39	620	12	42.8	1033
19	65	568	36	39.0	1603
49	57	1498	35	55.0	1611
45	115	1365	24	46.0	1613
55	58	1687	43	59.2	1854
59	60	1639	47	51.5	2160
94	85	2872	79	61.8	2305
128	201	3655	180	61.5	3503
96	133	2912	61	58.8	3571
131	108	3921	103	48.8	3741
127	155	3866	127	55.0	4026
252	362	7684	158	70.0	10343
409	347	12446	169	107.8	11732
464	392	14098	331	70.5	15414
510	865	15524	372	63.5	18854

9.14 จงใช้ข้อมูลจากข้อ 9.13 ที่ได้แก้ไขปัญหา multicollinearity แล้วตรวจสอบว่ามีข้อมูลใดบ้างที่มีความผิดปกติในตัวแปรตามและข้อมูลใดบ้างที่มีอิทธิพล

9.15 จากข้อ 5.1 จงหาว่ามีข้อมูลใดบ้างที่มีความผิดปกติในตัวแปรอิสระและในตัวแปรตามรวมถึงข้อมูลใดบ้างที่มีอิทธิพล

9.16 จากข้อ 5.10 จงหาว่ามีข้อมูลใดบ้างที่มีอิทธิพล ท่านจะจัดการกับข้อมูลที่มีปัญหานี้อย่างไร