บทที่ 6
การวิเคราะห์ความแปรปรวนร่วม
(Analysis of Covariance)

6.1 การวิเคราะห์ความแปรปรวนร่วมสำหรับ CRD (One-way ANOVA)

ใน CRD ความคลาดเคลื่อนจากการทดลอง (experimental errors) อาจเกิดขึ้นได้มาก เนื่องจากความไม่เป็นเนื้อเดียวกัน (homogeneous) ของหน่วยทดลอง (experimental units) วิธีการอันหนึ่งซึ่งจะขจัดหรือลดความคลาดเคลื่อนจากการทดลองโดยไม่ใช้วิธีการแยกกลุ่ม (Blocking) ของหน่วยทดลอง ก็โดยใช้วิธีการวิเคราะห์ความแปรปรวนร่วม (ANOCOV)

การแยกกลุ่ม (Blocking) ของหน่วยทดลองเป็นวิธีที่ดีสำหรับการลดความคลาดเคลื่อนจากการทดลอง ใน Blocking Designs เราแยกหน่วยทดลองออกเป็นกลุ่ม ๆ โดยที่หน่วยที่อยู่ในกลุ่มเดียวกันจะเหมือน ๆ กัน (homogeneous) แต่หน่วยที่อยู่ต่างกลุ่มกันจะต่างกัน ตัวอย่างเช่น ในการทดสอบปฏิกริยาต่อน้ำหอม 4 ชนิดที่จะใช้ในสบู่ชนิดหนึ่ง คนที่ถูกถามอาจถูกแบ่งออกเป็นกลุ่ม ๆ ตามอายุ ดังนั้นความแตกต่างในเรื่องความชอบไม่ชอบ ซึ่งเนื่องมาจากอายุจะถูกขจัดออกไปจาก experimental error

อย่างไรก็ดี บ่อยครั้งที่เราไม่สามารถจะแบ่งกลุ่มของหน่วยทดลองได้ ดังนั้นวิธีการวิเคราะห์ความแปรปรวนร่วมจะเป็นวิธีที่จะช่วยขจัด experimental errors ให้ พิจารณาตัวอย่างที่เราต้องการศึกษาว่าภาพยนตร์ 8 เรื่องจะช่วยส่งเสริมการท่องเที่ยวภายในประเทศได้ต่างกันหรือไม่ ก่อนให้ผู้ชมซึ่งเป็นหน่วยทดลองชมภาพยนตร์ ผู้ชมจะต้องตอบคำถามเกี่ยวกับทัศนคติในด้านการท่องเที่ยวภายในประเทศก่อน หลังจากนั้นก็ฉายภาพยนตร์ที่ยาว 5 นาทีให้ชม หลังจากชมภาพยนตร์แล้วผู้ชมจะถูกถามความเห็นเกี่ยวกับภาพยนตร์และความต้องการที่จะไปท่องเที่ยวและอื่น ๆ ทันที ในการทดลองอย่างนี้เราต้องการอย่างมากที่จะจัดกลุ่มผู้ชมตามความคุ้นเคยและทัศนคติต่อการท่องเที่ยวในประเทศก่อนที่จะกำหนดเรื่องของภาพยนตร์ให้ชม แต่อย่างไรก็ดีในที่นี้การจัดกลุ่มของผู้ชมก่อนเป็นไปไม่ได้ เพราะเราไม่รู้จักหน่วยทดลองมาก่อนที่จะนำหน่วยทดลองมาศึกษาทีละคน และเราจะไม่ได้ข้อมูลเกี่ยวกับทัศนคติเริ่มต้นของผู้ชม จนกว่าจะได้กำหนดภาพยนตร์ให้ชม

6.2 ตัวแบบของการวิเคราะห์ความแปรปรวนร่วม

(Analysis of Covariance Model)

$Y_{ij} = \mu + \alpha_i + \gamma(x_{ij} - \bar{x}) + E_{ij}$ หรือ $Y_{ij} = \mu_i + \gamma(x_{ij} - \bar{x}) + E_{ij}$

$i = 1, \ldots, t$ และ $j = 1, \ldots, n_i$

โดยที่ μ เป็น overall mean

α_i เป็น treatment effects โดยมีข้อจำกัดว่า $\sum_{i=1}^{t} n_i\alpha_i = 0$

γ เป็นสัมประสิทธิ์ของการถดถอย (regression coefficient) ซึ่งเท่ากันสำหรับทุก reatment

x_{ij} เป็น concomitant variable ซึ่งถือว่าเป็นค่าคงที่ (constant) ไม่ใช่ตัวแปรเชิงสุ่ม

$\bar{x}$ เป็น grand mean

E_{ij} คือ random error แต่ละตัวมีการกระจายเป็น $N(0, \sigma^2)$ และต่างก็เป็นอิสระต่อกัน

$\mu_i = \mu + \alpha_i$ เป็น mean ของ population ที่ i

ข้อสังเกต

1) ANOCOV เป็นเรื่องเกี่ยวกับตัวแปร 2 ตัวเป็นอย่างน้อย และเราไม่สามารถควบคุมตัวแปรอิสระ (independent variable) ได้ เช่น ถ้า $Y = f(x)$ เราควรใช้ ANOCOV ถ้าเราควบคุม x ไม่ได้

2) ตัวแบบของ ANOCOV มีเทอม $\gamma(x_{ij} - \bar{x})$ ซึ่งแสดงความสัมพันธ์ของ X และ Y เหมือนในตัวแบบการถดถอย (Regression model) เพิ่มขึ้นมาจากตัวแบบของ ANOVA ดังนั้นอาจกล่าวโดยทั่วไปว่า ANOCOV เป็นการรวมเทคนิคทางสถิติ 2 อย่างเข้าด้วยกันคือ การวิเคราะห์การถดถอย (Regression analysis) และการวิเคราะห์ความแปรปรวน (Analysis of variance)

6.3 ข้อสมมุติในการวิเคราะห์ความแปรปรวนร่วม

ข้อสมมุติที่สำคัญที่ทำให้การวิเคราะห์ความแปรปรวนร่วมถูกต้องมีดังนี้

1) X เป็นค่าคงที่ เราสามารถวัด X ได้โดยไม่มีความผิดพลาด

2) Regression ของ Y on X หลังจากการที่เราแยกความแปรปรวนเนื่องจากวิธีการ (treatment) ออกแล้ว จะเป็น linear regression ที่ปราศจากอิทธิพลของวิธีการ (treatment effect)

3) Error มีการกระจายเป็น Normal และเป็นอิสระจากกันและกัน มีค่าเฉลี่ยเท่ากับศูนย์ และ variance เท่ากันหมด

6.4 การสร้าง ANOCOV table เพื่อทดสอบสมมติฐาน

เราต้องการทดสอบสมมติฐาน $H_0 : \alpha_i = 0, \forall_i = 1, \ldots, t$, H_1 : at least one $\alpha_i \neq 0$ หรือทดสอบ $H_0 : \mu_1 = \ldots = \mu_t$, H_1 : มี Mean อย่างน้อย 1 คู่ไม่เท่ากัน

โครงสร้างของข้อมูล

	วิธีการที่ 1		วิธีการที่ 2		...	วิธีการที่ t			
	x	y	x	y		x	y		
	x_{11}	y_{11}	x_{21}	y_{21}		x_{t1}	y_{t1}		
	x_{12}	y_{12}	x_{22}	y_{22}		x_{t2}	y_{t2}		
	⋮	⋮	⋮	⋮		⋮	⋮		
	x_{1n_1}	y_{1n_1}	x_{2n_2}	y_{2n_2}		x_{tn_t}	y_{tn_t}		
Total	$x_{1.}$	$y_{1.}$	$x_{2.}$	$y_{2.}$		$x_{t.}$	$y_{t.}$	G_x	G_y
Mean	$\bar{x}_1$	$\bar{y}_1$	$\bar{x}_2$	$\bar{y}_2$		$\bar{x}_t$	$\bar{y}_t$	$\bar{x}$	$\bar{y}$
sample size	n_1		n_2			n_t		$N = \sum_{i=1}^{t} n_i$	

โดยที่ $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$, $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$

$\bar{x}_i = \frac{x_{i.}}{n_i}$, $\bar{y}_i = \frac{y_{i.}}{n_i}$

$G_x = \sum_i \sum_j x_{ij} = \sum_i x_{i.}$, $G_y = \sum_i \sum_j y_{ij} = \sum_i y_{i.}$

$\bar{x} = G_x/N$, $\bar{y} = G_y/N$

ในการสร้าง ANOCOV table คำนวณค่าต่าง ๆ ต่อไปนี้คือ

สำหรับ y; $(CF)_y = \frac{G_y^2}{N}$

$$SST_y = \sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j y_{ij}^2 - (CF)_y$$

$$SStr_y = \sum_i n_i (\bar{y}_i - \bar{y})^2 = \sum_i \frac{y_{i.}^2}{n_i} - (CF)_y$$

$$SSE_y = SST_y - SStr_y$$

สำหรับ x; $(CF)_x = \frac{G_x^2}{N}$

$$SST_x = \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j x_{ij}^2 - (CF)_x$$

$$SStr_x = \sum_i n_i (\bar{x}_i - \bar{x})^2 = \sum_i \frac{x_{i.}^2}{n_i} - (CF)_x$$

$$SSE_x = SST_x - SStr_x$$

สำหรับ sum of products (SP); (ค่าอาจเป็นลบได้)

$$(CF)_{xy} = \frac{G_x G_y}{N}$$

$$SPT = \sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y}) = \sum_i \sum_j x_{ij} y_{ij} - (CF)_{xy}$$

$$SPtr = \sum_i n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = \sum_i \frac{x_{i.} y_{i.}}{n_i} - (CF)_{xy}$$

$$SPE = SPT - SPtr$$

เนื่องจาก SST คือ ความแปรปรวนที่เนื่องมาจากความแตกต่างของวิธีการแล้ว ยังรวมเอาความแปรปรวนที่เนื่องมาจากการถดถอย (อิทธิพลของ X) ด้วย ดังนั้นก่อนที่จะทำการทดสอบความแตกต่างของวิธีการ เราต้องขจัดอิทธิพลที่เนื่องมาจาก X เสียก่อน

พิจารณาการวิเคราะห์การถดถอย (หัวข้อ 5.3)

$$SSE = SST - SSR$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - b_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{[\Sigma (x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma (x_i - \bar{x})^2}$$

$$\because b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

นั่นคือ SSE คือ SST ที่ขจัดส่วนของความแปรปรวนที่เนื่องมาจากอิทธิพลของ X ออกแล้ว ถ้าเขียน SSE ในเทอมของ ANOCOV (โดยที่เราใช้ x_{ij} และ y_{ij} แทน x_i, y_i) จะได้

$$SSE = SST_y - \frac{(SPT)^2}{SST_x}$$

ดังนั้น ใน ANOCOV ความแปรปรวนทั้งหมดของ Y ที่ขจัดส่วนของความแปรปรวนที่เนื่องมาจากอิทธิพลของ X แล้วคือ

$$SST\ (adjusted) = SST_y - \frac{(SPT)^2}{SST_x} \quad , \quad d.f. = N - 2$$

โดยที่ SST(adj.) อ่านว่า adjusted total sum of squares for y

ทำนองเดียวกัน

$$SSE(adj.) = SSE_y - \frac{(SPE)^2}{SSE_x} \quad , \quad d.f. = N - t - 1$$

SSE(adj.) อ่านว่า adjusted error sum of squares for y

ถ้า SStr(adj.) = adjusted treatment sum of squares for y

ดังนั้น SStr(adj.) = SST(adj.) − SSE(adj.) , d.f. = t − 1

$$MStr(adj.) = \frac{SStr(adj.)}{t - 1}$$

$$MSE(adi.) = \frac{SSE(adj.)}{N - t - 1}$$

ANOCOV

S.V.	d.f.	SS และ SP			adj d.f.	y adjusted for x		
		x	y	xy		adj. SS	adj. MS	f.
Treatments	t − 1	$SStr_x$	$SStr_y$	SPtr	t − 1	SStr(adj.)	MStr(adj.)	$\frac{MStr\ (adj.)}{MSE\ (adj.)}$
Error	N − t	SSE_x	SSE_y	SPE	N − t − 1	SSE(adj.)	MSE(adj.)	
Total	N − 1	SST_x	SST_y	SPT	N − 2	SST(adj.)		

ในการทดสอบ $H_0 : \alpha_i = 0, \forall_i = 1, \ldots, t$

H_1: at least one $\alpha_i \neq 0$

ถ้า H_0 จริง $f_c = \frac{MStr(adj.)}{MSE(adj.)}$ จะเป็นค่าของตัวสถิติ F ซึ่งมีการกระจาย

เป็น F - distribution ที่มี d.f. = (t − 1, N − t − 1)

ตัวอย่างที่ 6.1 ในการทดลองเพื่อเปรียบเทียบอิทธิพลของยา 4 ชนิดในการทำให้กล้ามเนื้อส่วนหนึ่งหดตัว หนู 4 กลุ่ม ถูกให้ยาอย่างสุ่ม หลังจากนั้น 12 วัน ฆ่าหนู และชั่งน้ำหนักของกล้ามเนื้อของหนู (Y) การหดตัวของกล้ามเนื้อควรจะถูกวัดด้วยน้ำหนักที่หายไปของกล้ามเนื้อส่วนนั้น แต่เราไม่สามารถจะได้น้ำหนักเริ่มต้นของกล้ามเนื้อส่วนนั้นถ้าเราไม่ฆ่าหนู ดังนั้นเราจึงใช้น้ำหนักเริ่มต้นของหนู (X) แทน เพราะเราเชื่อว่าน้ำหนักเริ่มต้นของหนูนั้นจะต้องเกี่ยวข้องกับน้ำหนักเริ่มต้นของกล้ามเนื้อส่วนนั้นด้วย

ตารางแสดงค่าของ x = น้ำหนักเริ่มต้นของหนู และ
y = น้ำหนักของกล้ามเนื้อส่วนที่สนใจภายหลังได้รับยาแล้ว .

	ยา A		ยา B		ยา C		ยา D	
	x	y	x	y	x	y	x	y
	198	.34	233	.41	204	.57	186	.81
	175	.43	250	.87	234	.80	286	1.01
	199	.41	289	.91	211	.69	245	.97
	224	.48	255	.87	214	.84	215	.87
Total	$x_{1.} = 796$	$y_{1.} = 1.66$	1027	3.06	863	2.9	932	3.66

$G_x = 796 + 1027 + 863 + 932 = 3618$

$G_y = 1.66 + 3.06 + 2.9 + 3.66 = 11.28$

$N = 16$

Model : $y_{ij} = \mu + \alpha_i + \gamma(x_{ij} - \bar{x}) + E_{ij}$, $i = 1, \ldots, 4$ และ $j = 1, \ldots, 4$

ต้องการทดสอบ $H_0 : \alpha_i = 0, \forall i = 1, \ldots, 4$

คำนวณค่าต่าง ๆ ดังนี้

สำหรับ y ; $(CF)_y = \dfrac{(11.28)^2}{16} = 7.9524$

$$SST_y = \{(.34)^2 + (.43)^2 + \ldots + (.87)^2\} - (CF)_y = .7772$$

$$SStr_y = \frac{(1.66^2 + 3.06^2 + 2.9^2 + 3.66^2)}{4} - (CF)_y = .5288$$

$$SSE_y = .7772 - .5288 = .2484$$

สำหรับ x ; $(CF)_x = \frac{(3618)^2}{16} = 818120.25$

$$SST_x = (198^2 + 175^2 + \ldots + 215^2) - (CF)_x = 16151.75$$

$$SStr_x = \frac{(796^2 + \ldots + 932^2)}{4} - (CF)_x = 7314.25$$

$$SSE_x = 16151.75 - 7314.25 = 8837.5$$

สำหรับ Sum of products;

$$(CF)_{xy} = \frac{(3618)\,(11.28)}{16} = 2550.69$$

$$SPT = \{ (198)\,(.34) + \ldots + (215)\,(.87) \} - (CF)_{xy} = 71.91$$

$$SPtr = \frac{(796)(1.66) + (1027)(3.06) + (863)(2.9) + (932)(3.66)}{4} - (CF)_{xy}$$

$$= 43.76$$

$$SPE = 71.91 - 43.76 = 28.15$$

$$SST\ (adj.) = SST_y - \frac{(SPT)^2}{SST_x} = .7772 - \frac{(71.91)^2}{16151.75} = .4570$$

$$SSE\ (adj.) = SSE_y - \frac{(SPE)^2}{SSE_x} = .2484 - \frac{(28.15)^2}{8837.5} = .1587$$

$$\therefore SStr\ (adj.) = .4570 - .1587 = .2982$$

ANOCOV

S.V.	d.f.	SS และ SP			adj d.f.	y adjusted for x		
		SS_x	SS_y	SP		SS (adj.)	MS.(adj.)	f_c
Treatments	3	7314.25	.5288	46.76	3	.2982	.0994	6.89
Error	12	8837.5	.2484	28.15	11	.1587	.0144	
Total	15	16151.75	.7772	71.91	14	.4570		

1) $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

2) $H_1 : \alpha_i$ อย่างน้อย 1 ตัว $\neq 0$

3) $\alpha = .05$

4) CR : $F > f_{(3, 11), .05} = 3.59$

5) $f_c = \dfrac{\text{MStr (adj.)}}{\text{MSE (adj.)}} = \dfrac{.0994}{.0144} = 6.89$

6) $\because f_c = 6.89 > 3.59$ เราปฏิเสธ H_0 ที่ $\alpha = .05$ (ในที่นี่เราสามารถปฏิเสธ H_0 ที่ $\alpha = .01$ ด้วย เพราะ $f_c > f_{(3, 11), .01} = 6.22$) นั่นคือ α_i ไม่เท่ากับศูนย์ทุกตัว หรือสรุปได้ว่า population means ไม่เท่ากันหมด

หมายเหตุ

ถ้าใช้ ANOVA, MSE $= \dfrac{.2484}{12} = .0207$

แต่ถ้าใช้ ANOCOV, MSE (adj.) $= \dfrac{.1587}{11} = .0144$ จะเห็นว่า การใช้ ANOCOV ทำให้ Mean square error ลดลง

แบบฝึกหัด

6.1 ในการทดลองเพื่อเปรียบเทียบอิทธิพลของปริมาณปุ๋ย 3 ระดับ ต่อผลผลิต (Y) ของพืชชนิดหนึ่ง ได้จดบันทึกจำนวนต้นพืชในแปลงหนึ่ง ๆ (X) ไว้ด้วย จงวิเคราะห์ข้อมูลต่อไปนี้

ปุ๋ย					
ระดับที่ 1		ระดับที่ 2		ระดับที่ 3	
x	y	x	y	x	y
65	30	34	46	26	52
61	27	31	52	23	59
47	43	30	48	48	46
52	27	35	45	32	45
49	51	49	51	25	44

6.2 ในการทดลองเพื่อดูความแตกต่างของอาหาร 4 ชนิดต่อน้ำหนักที่เพิ่มขึ้น น้ำหนักเริ่มต้น (X)และน้ำหนักที่เพิ่มขึ้นเมื่อหมดระยะเวลาของการทดลอง (Y) เป็นดังนี้ จงวิเคราะห์ข้อมูลนี้

อาหาร							
A		B		C		D	
x	y	x	y	x	y	x	y
38	9.52	39	8.51	48	9.11	48	9.75
35	8.21	38	9.95	37	8.5	28	8.66
41	9.32	46	8.43	42	8.9	33	7.63
48	4.56	40	8.86	42	9.51	50	9.37