เฉลยข้อสอบชุดที่ 1

<u>ข้อสอบชุดนี้มี 5 ข้อ 100 คะแนน</u>

ข้อ 1. จงเติมคำหรือข้อความให้ถูกต้องและได้ความ

1.1 เมื่อแบบจำลอง $Y_{n\times1} = X_{n\times p}\beta_{p\times1} + e_{n\times1}$ มี rank(X) = p เราเรียกแบบจำลองนี้ว่า General linear hypothesis model of full rank (แบบจำลอง a)

แต่ถ้า rank(X) < p เราจะเรียกแบบจำลอง (แบบจำลอง b) ว่า

<u>General linear hypothesis model of less than full rank</u>

1.2 จากข้อ 1.1 แบบจำลอง (a) หรือ (b) ที่คือ Experimental Design Model

<u>(b)</u>

1.3 Regression Model เป็นแบบจำลอง (a) หรือแบบจำลอง (b) ใน 1.1 <u>(a)</u>

1.4 จากข้อ 1.1 สำหรับแบบจำลอง (a) จงเขียน normal equations

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

ดังนั้น least square estimator ของ β คือ $\underline{\hat{\beta} = (X'X)^{-1}X'Y}$

1.5 จงเขียน

(a) Cubic polynomial model: $\underline{y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3 + e}$

และ (b) Quartic polynomial model: $\underline{y = \lambda_0 + \lambda_1 x + \ldots + \lambda_4 x^4 + e}$

1.6 จาก Regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$

R^2 (Coefficient of multiple determination) = $SS(\beta_1,\beta_2,\beta_3 | \beta_0)/SST$

ใช้อธิบาย <u>ส่วนของความแปรปรวนของ Y ซึ่งเป็นผลมาจากอิทธิพลของ x_1, x_2 และ x_3</u>

1.7 จาก 1.6 $R^2 = \sqrt{R}$ คือสหสัมพันธ์ชนิดหนึ่งซึ่งมีชื่อว่า **Multiple correlation coefficient**

1.8 จาก 1.6 ถ้า partial $f_c(X_3 | X_1, X_2)$ ไม่มีนัยสำคัญ (n.s.) เราสรุปผลว่า **ไม่ควรเพิ่ม X_3 เข้าไปในแบบจำลองซึ่งมี X_1 และ X_2 อยู่แล้ว**

1.9 ใน (a x b) factorial design นั้น ถ้าระดับของปัจจัย A ถูกสุ่มมาจากระดับทั้งหมดของ A และระดับของปัจจัย B ถูกกำหนดขึ้นตามความสนใจ เราเรียกแบบจำลองของการวิเคราะห์ว่า **Mixed model**

1.10 ใน (5 x 5) Greaco latin square design (1 หน่วยทดลอง/cell) เราสามารถทดสอบอิทธิพลหลักได้เป็นจำนวน 4 อิทธิพล แต่เราใช้หน่วยทดลองสำหรับ design นี้เพียง $5^2 = 25$ หน่วยทดลองเท่านั้น

ข้อ 2. จากชายผิวขาว 32 คน ซึ่งมีอายุเกิน 40 ปี ได้บันทึกข้อมูลต่อไปนี้ไว้คือ

(Y) SBP = ความดันโลหิต

(X_1) AGE = อายุ

(X_2) SMK = ประวัติการสูบบุหรี่ (=1 ถ้ายังสูบบุหรี่อยู่หรือเคยสูบบุหรี่)

(X_3) QUET= ขนาดของร่างกาย (QUET = "Quetelet index" คือดรรชนีวัดขนาดของร่างกาย โดยที่ QUET = 100(weight/height^2))

คนที่	Y	X_1	X_2	X_3	คนที่	Y	X_1	X_2	X_3
1	135	45	0	2.876	17	145	49	1	3.360
2	122	41	0	3.251	18	142	46	1	3.024
3	130	49	0	3.100	19	135	57	0	3.171
4	148	52	0	3.768	20	142	56	0	3.401
5	146	54	1	2.979	21	150	56	1	3.628
6	129	47	1	2.790	22	144	58	0	3.751
7	162	60	1	3.668	23	137	53	0	3.296
8	160	48	1	3.612	24	132	50	0	3.210
9	144	44	1	2.368	25	149	54	1	3.301
10	180	64	1	4.637	26	132	48	1	3.017
11	166	59	1	3.877	27	120	43	0	2.789
12	138	51	1	4.032	28	126	43	1	2.956
13	152	64	0	4.116	29	161	63	0	3.800
14	138	56	0	3.673	30	170	63	1	4.132
15	140	54	1	3.562	31	152	62	0	3.962
16	134	50	1	2.998	32	164	65	0	4.010

ในการศึกษาถึงความสัมพันธ์ในรูปการถดถอยแบบพหุคูณของ SBP(Y) กับ AGE(X_1), SMK(X_2) และ QUET(X_3) ได้พิจารณาแบบจำลองการถดถอยแบบพหุคูณ 3 แบบจำลองคือ

1) y บน x_1

2) y บน x_1, x_2

3) y บน x_1, x_2, x_3

จาก Computer Printer **จงเติมตาราง และ ตอบปัญหาต่อไปนี้**

(ให้ใช้ทศนิยม 1 ตำแหน่ง)

Least Squares estimates

แบบจำลอง	ตัวแปรอิสระในแบบจำลอง	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$S_{\hat{\beta}_1}$	$S_{\hat{\beta}_2}$	$S_{\hat{\beta}_3}$
1	AGE (X_1)	59.1	1.6	-	-	0.2	-	-
2	AGE (X_1), SMK (X_2)	48.0	1.7	10.3	-	0.2	2.8	-
3	AGE (X_1), SMK (X_2), QUET(X_3)	45.1	1.2	9.9	8.6	0.3	2.7	4.5

ตาราง ANOVA

แบบจำลอง 1		
SV	df	SS
Regression (X_1)	1	3861.6
Error	30	2564.3

แบบจำลอง 2		
SV	df	SS
Regression (X_1, X_2)	2	4689.1
Error	29	1736.3

แบบจำลอง 3		
SV	df	SS
Regression (X_1, X_2, X_3)	3	4889.8
Error	28	1536.1

1) แบบจำลอง 3: $\widehat{SBP}$ = 45.1 + 1.2(AGE) + 9.9(SMK) + 8.6(QUET)

1.1 ทำนาย SBP ของชายอายุ 50 ปี สูบบุหรี่ และ ขนาดของร่างกาย = 3.5

$\widehat{SBP}$ = 45.1 + 1.2(50) + 9.9(1) + 8.6(3.5) = 145.1 ...(1)

1.2 ทำนาย SBP ของชายอายุ 50 ปี ไม่สูบบุหรี่ และ ขนาดของร่างกาย = 3.5

$\widehat{SBP}$ = 45.1 + 1.2(50) + 9.9(0) + 8.6(3.5) = 135.2

1.3 สำหรับชายอายุ 50 ปี สูบบุหรี่ จงประมาณความแตกต่างของ SBP เมื่อขนาดของร่างกายเพิ่มจาก 3.0 เป็น 3.5

$\widehat{SBP}$ = 45.1 + 1.2(50) + 9.9(1) + 8.6(3.0) = 140.8 ...(2)

(1) - (2) = 4.3

นั่นคือ SBP เพิ่มขึ้นประมาณ 4.3 เมื่อขนาดของร่างกายเพิ่มจาก 3.0 เป็น 3.5

2) R^2 ของแบบจำลองทั้ง 3 คือ

Model 1: R^2 = 0.60094, Model 2: R^2 = 0.72980

Model 3: R^2 = 0.76095

3) จงทำ overall F-test ของแต่ละแบบจำลอง ระบุ H_o ที่ทดสอบให้ชัดเจนด้วย

Model 1: $H_o: \beta_1=0$, $f_c=45.17692^{**}$, Signif F = .0000 < .01

เราปฏิเสธ H_o ที่ α = .01

[หรือปฏิเสธ H_o ที่ α = .05 เพราะ $f_c > f_{(1,30),.05}=4.17$]

Model 2: $H_o: \beta_1=\beta_2=0$, $f_c=39.16433^{**}$, Signif F = .0000 < .01

เราปฏิเสธ H_o ที่ α = .01

[หรือปฏิเสธ H_o ที่ α = .05 เพราะ $f_c > f_{(2,29),.05}=3.33$]

Model 3: $H_o: \beta_1=\beta_2=\beta_3=0$, $f_c=29.70972^{**}$, Signif F = .0000 < .01

เราปฏิเสธ H_o ที่ α = .01

[หรือปฏิเสธ H_o ที่ α = .05 เพราะ $f_c > f_{(3,28),.05}=2.95$]

4) จงทำ partial F-test โดยใช้ α = .05 สำหรับการ

4.1 เพิ่ม SMK เข้าไปในแบบจำลองที่มี AGE อยู่แล้ว

4.2 เพิ่ม QUET เข้าไปในแบบจำลองที่มี AGE และ SMK อยู่แล้ว

ข้อนี้นักศึกษาอาจทำ t-test ซึ่งสมนัยกับ partial F-test ก็ได้

4.1) Partial T(SMK | AGE) = 3.719, sig T = 0.0009 < .05

(จาก printout หน้า 13)

ดังนั้นปฏิเสธ H_o: $\beta_2 = 0$ ที่ α = .05

[Partial $f_c = (3.719)^2 = 13.8310$]

4.2 Partial T(QUET | SMK,AGE) = 1.91, sig T = 0.0664 > .05

(จา printput หน้า 15)

ดังนั้นไม่ปฏิเสธ H_o: $\beta_3 = 0$ ที่ α = .05

[Partial $f_c = (1.91)^2 = 3.6481$]

ข้อ 3. จากโจทย์และ Computer printout ในข้อ 2. จงอธิบายวิธีเลือกแบบจำลองการถดถอยที่ดีที่สุด (Best Regression model) สำหรับการทำนาย Y (SBP) **เลือกอธิบายอย่างละเอียด 1 วิธี** จาก 3 วิธีที่กำหนดให้ คือ

1) วิธี Forward selection

2) วิธี Backward elimination

3) วิธี Stepwise

<u>หมายเหตุ</u> ให้ใช้ค่าจาก Computer printout ประกอบคำอธิบายทุกขั้นตอน

<u>เฉลย</u> <u>ดู</u>เฉลยแบบฝึกหัดข้อ 6.6

ข้อ 4.1 ตารางต่อไปนี้แสดงระยะเวลารอคอยเฉลี่ย (นาที) ที่คนไข้ 16 คนต้องรอหมอ คนไข้ถูกจำแนกโดย ชนิดของการฝึกหัดของหมอ (Type of practice: TYPRAC) และชนิดของหมอ (Type of physician: PHYSTY)

	ชนิดของการฝึกหัด	
ชนิดของหมอ	แบบกลุ่ม (GROUP)	แบบเดี่ยว (SOLO)
รักษาโรคทั่วไป(GP)	15,20,25,20	20,25,30,25
เฉพาะทาง(SPEC)	30,25,30,35	25,20,30,30

กำหนดบางส่วนของตาราง ANOVA ให้ **จงเติมตาราง ANOVA ให้สมบูรณ์ และจงทดสอบที่ α = .05 สำหรับอิทธิพลหลัก และอิทธิพลร่วม**
โดยใช้ **Random effect model**

ANOVA

SV	df	SS	MS	f
PHYSTY	1	126.56	(1) 126.56	f_1=(1)/(3)=1.65(n.s.)
TYPRAC	1	1.5625	(2) 1.5625	f_2=(2)/(3)=.026(n.s.)
PHYSTY x TYPRAC	1	76.563	(3) 76.563	f_3=(3)/(4)=4.28(n.s.)
Error	12	218.7525	(4) 18.23	
Total	15	123.438		

<u>ทดสอบอิทธิพลร่วม</u> $f_3 < f_{(1,12),.05} = 4.75$

สรุปว่าไม่มีอิทธิพลร่วมระหว่าง PHYSTY และ TYPRAC

<u>ทดสอบอิทธิพลหลัก PHYSTY</u> $f_1 < f_{(1,1),.05} = 161.4$

สรุปว่าชนิดของหมอไม่มีอิทธิพลต่อระยะเวลารอคอย

<u>TYPRAC</u> $f_2 < f_{(1,1),.05} = 161.4$

สรุปว่าชนิดของการฝึกหัดไม่มีอิทธิพลต่อระยะเวลารอคอย

ข้อ 4.2 ที่ปรึกษาเกี่ยวกับ MIS (Management Information System) ได้ทำ small-scale study เกี่ยวกับรายงานสรุปรายวัน 5 แบบ (A = มีรายละเอียดมากที่สุด, B, C, D, E = มีรายละเอียดน้อยที่สุด) เขาใช้ผู้จัดการฝ่ายขาย (Sales executive) 5 คน แต่ละคนจะได้รับรายงานสรุปรายวันแบบหนึ่งโดยตลอดในเดือนหนึ่ง ๆ แล้วจะต้องให้คะแนนว่ารายงานนั้นช่วยการตัดสินใจมากน้อยเพียงใดตาม 25-point scale (0 = ไม่ช่วยเลย,..., 25 = ช่วยมากที่สุด) ในช่วงเวลา 5 เดือน ผู้จัดการฝ่ายขายแต่ละคนจะได้รับรายงานแบบหนึ่งในเดือนหนึ่ง ๆ ผลปรากฏในตารางต่อไปนี้

ผู้จัดการ	เดือน				
	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.
นาย ก	D 21	A 8	C 17	B 9	E 16
นาย ข	A 5	E 10	B 3	C 12	D 15
นาย ค	C 20	B 10	E 15	D 21	A 12
นาย ง	B 4	D 15	A 3	E 9	C 10
นาย จ	E 17	C 16	D 20	A 8	B 11

กำหนดบางส่วนของตาราง ANOVA ให้ **จงเติมตาราง ANOVA ให้สมบูรณ์ และจงทดสอบที่ α = .025 ว่า แบบของรายงานมีอิทธิพลต่างกันหรือไม่**

SV	df	SS	MS	f
ผู้จัดการ	4	233.04	**58.26**	f_1 = 62.87*
เดือน	4	12.24	**3.06**	f_2 = 3.30*
รายงาน	4	**478.64**	**119.66**	f_3 = 129.13*
Error	**12**	11.12	**0.9267**	
Total	**24**	735.04		

$f_3 > f_{(4,12),.05} = 3.26$ สรุปว่าแบบของรายงานมีอิทธิพลต่างกัน

ข้อ 5.1 จากแบบจำลอง $y_{ij} = \mu_i + e_{ij}$, $i = 1,\ldots,3;\ j = 1,\ldots, n_i$

ค่าสังเกต:

	วิธีการ			
	1	2	3	
	y_{11} y_{12} y_{13} y_{14}	y_{21} y_{22}	y_{31} y_{32} y_{33}	
n_i	4	2	3	$n = 9$
T_i	T_1	T_2	T_3	G

ถ้าแบบจำลองในรูปเมตริกซ์ คือ $Y = X\beta + e$

จงเขียนสิ่งต่อไปนี้ (โดยแสดงสมาชิกให้ชัดเจน และบอกขนาดของเมตริกซ์ หรือเวคเตอร์ด้วย) Y, X, β, e, $X'X$, $X'Y$, $Y'Y$ และ $e'e$

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix}_{9\times1}, \quad X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}_{9\times3}, \quad \beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}_{3\times1}, \quad e = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \\ e_{33} \end{bmatrix}_{9\times1}$$

$$X'X = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}_{3x3} , \quad X'Y = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}_{3x1}$$

$$Y'Y = \sum_{i=1}^{3} \sum_{j=1}^{n} y_{ij}^2 \text{ และ } e'e = \sum_{i=1}^{3} \sum_{j=1}^{n} e_{ij}^2$$

ข้อ 5.2 จากข้อมูลเกี่ยวกับปริมาณน้ำฝนในปีต่าง ๆ สมมุติว่าเราต้องการหา polynomial ที่ fit กับข้อมูลดังต่อไปนี้คือ

ปีที่ i	1	2	3	4	5	6	7	8
y:ปริมาณน้ำฝน	30.2	32.2	35.1	34.2	39.1	41.3	36.1	30.1

ปีที่ i	9	10	11	12
y:ปริมาณน้ำฝน	30.5	26.1	24.8	28.2

$$\sum_{i=1}^{12} y_i = 387.9 , \quad \sum_{i=1}^{12} y_i^2 = 12815.99$$

จากตาราง ANOVA ที่กำหนดให้

1. จงทดสอบสมมติฐานที่ α = .02 ว่า linear term, quadratic term, cubic term และ quartic term มีนัยสำคัญหรือไม่ (ให้แสดงการทดสอบให้ชัดเจนประกอบการสรุปผล

Linear model: $y = \alpha_0 + \alpha_1 x + e$

ทดสอบ H_0: $\alpha_1 = 0$, $f_c = 3.48$(n.s.) , Pr > 2% นั่นคือ $\alpha_1 = 0$

Quadratic model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$

ทดสอบ H_0: $\beta_2 = 0$, $f_c = 9.21^*$, Pr < 2% นั่นคือ $\beta_2 \neq 0$

Cubic model: $y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3 + e$

ทดสอบ H_0: $\gamma_3 = 0$, $f_c = 4.88$(n.s.) , Pr > 2% นั่นคือ $\gamma_3 = 0$

Quartic model: $y = \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3 + \lambda_4 x^4 + e$

ทดสอบ H_0: $\lambda_4 = 0$, $f_c = 8.40$(n.s.) , Pr > 2% นั่นคือ $\lambda_4 = 0$

2. จากการทดสอบสมมติฐานในข้อ 1. สรุปว่า Quadratic model fit กับข้อมูลข้างต้นดีที่สุด หรือ polynomial degree 2 fit กับข้อมูลข้างต้นดีที่สุด

ANOVA

S.V.	df.	SS	MS	f_c	Pr
Total (uncorrected)	12	12815.99			
Reduction for mean ($n\bar{y}^2$)	1	12538.87			
Linear	1	71.548	71.548	3.48	>2%
Error	10	205.576	20.558		
Quadratic	1	103.963	103.963	9.21	<2%
Error	9	101.613	11.290		
Cubic	1	38.484	38.484	4.88	>2%
Error	8	63.130	7.891		
Quartic	1	34.432	34.432	8.40	>2%
Error	7	28.691	4.100		