แผนบริหารการสอนประจำบทที่ 4

วิชา 4114601 การวิเคราะห์สถิติขั้นสูงด้วยโปรแกรมสำเร็จรูปทางสถิติ
แผนบริหารการสอน บทที่ 4 การวิเคราะห์การถดถอยโลจิสติค เวลา 8 ชั่วโมง

สาระสำคัญ

การวิเคราะห์การถดถอยโลจิสติค จะเหมือนกับการวิเคราะห์ถดถอยโดยทั่วไป โดยจะเป็นการหาความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระ ซึ่งเมื่อได้แบบแผนความสัมพันธ์ (สมการพยากรณ์) แล้ว สามารถนำแบบแผนดังกล่าวไปใช้ประมาณค่าตัวแปรตามได้ โดยข้อแตกต่างระหว่างการวิเคราะห์การถดถอยโลจิสติค และ การวิเคราะห์ถดถอยโดยทั่วไป ก็คือ ตัวแปรตาม (Y) เป็นตัวแปรเชิงกลุ่ม แต่ในการวิเคราะห์ถดถอยโดยทั่วไปตัวแปรตาม (Y) จะต้องเป็นตัวแปรเชิงปริมาณ

ผลการเรียนรู้ที่คาดหวัง

นักศึกษาสามารถมองเห็นข้อแตกต่างระหว่างการวิเคราะห์ถดถอยโดยทั่วไปและการวิเคราะห์การถดถอยโลจิสติค และสามารถวิเคราะห์ผลพร้อมทั้งสรุปผลที่ได้จากการวิเคราะห์ได้อย่างถูกต้อง

จุดประสงค์การเรียนรู้

1. เข้าใจหลักการของการวิเคราะห์การถดถอยโลจิสติค
2. เปรียบเทียบความแตกต่างระหว่างเทคนิค การวิเคราะห์การถดถอยโลจิสติค กับการวิเคราะห์ถดถอยโดยทั่วไปได้
3. สามาถประยุกต์ใช้เทคนิค การวิเคราะห์การถดถอยโลจิสติค ในงานด้านต่างๆได้
4. สามารถสรุปผลที่ได้จากการวิเคราะห์ด้วยเทคนิค การวิเคราะห์การถดถอยโลจิสติคได้อย่างถูกต้อง

กิจกรรมการเรียนการสอน

1. นำเสนอ PowerPoint เนื้อหาเกี่ยวกับการวิเคราะห์การถดถอยโลจิสติคจากเอกสารประกอบการสอนการวิเคราะห์สถิติขั้นสูงด้วยโปรแกรมสำเร็จรูปทางสถิติ
2. ให้นักศึกษาค้นคว้าข้อมูลงานวิจัยที่เลือกใช้การวิเคราะห์การถดถอยโลจิสติคในงานวิจัย และนำมาอภิปรายในชั้นเรียน
3. ให้นักศึกษาทำแบบฝึกหัดท้ายบท

สื่อการเรียนรู้

1. ตัวอย่างงานวิจัย
2. PowerPoint

3. เอกสารประกอบการสอนรายวิชา 4114601 การวิเคราะห์สถิติขั้นสูงด้วยโปรแกรมสำเร็จรูปทางสถิติ

การวัดและประเมินผล

1. ประเมินผลจากแบบฝึกหัด
2. ประเมินผลจากการถาม – ตอบในชั้นเรียน
3. ประเมินผลจากการทดสอบย่อยรายบท

บทที่ 4
การวิเคราะห์การถดถอยโลจิสติค
(Logistic Regression)

ตัวแบบการถดถอยเชิงเส้นที่ใช้วิธีกำลังสองน้อยที่สุด (least square method) ในการประมาณค่าตัวแบบ มีข้อตกลงเบื้องต้นหลายประการด้วยกัน เช่น ตัวแปรต้นและตัวแปรตามต้องเป็นตัวแปรแบบต่อเนื่อง ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามต้องเป็นเชิงเส้นตรง การแจกแจงของความคลาดเคลื่อนต้องเป็นการแจกแจงแบบปกติ และมีค่าความแปรปรวนคงที่ ซึ่งจากข้อตกลงเบื้องต้นที่กล่าวมานี้ จะเห็นว่าข้อจำกัดที่สำคัญที่สุดข้อหนึ่งเกี่ยวกับตัวแปรในการถดถอยเชิงเส้น คือตัวแปรตามจะต้องเป็นแบบต่อเนื่อง (continuous variable) เท่านั้น ซึ่งบ่อยครั้งที่จะพบว่าในสถานการณ์ต่างๆข้อมูลที่นำมาวิเคราะห์มักจะเป็นตัวแปรตามแบบไม่ต่อเนื่องโดยเฉพาะในทางการแพทย์จะพบสถานการณ์เช่นนี้เสมอ เช่น การเกิดโรค กับ การไม่เกิดโรค , คนไข้รอดชีวิต กับ คนไข้เสียชีวิต , หรือในด้านอื่นๆ เช่น นักศึกษาที่สอบผ่าน กับ นักศึกษาที่สอบไม่ผ่าน เป็นต้น หรือบางครั้งตัวแปรเชิงกลุ่มอาจจะมีค่าสามค่าหรือมากกว่า เช่น นักศึกษาเลือกเรียนวิชาเอกสถิติประยุกต์ วิชาเอกคณิตศาสตร์ หรือวิชาเอกภาษาอังกฤษ เป็นต้น

1. ประเภทของการวิเคราะห์การถดถอยโลจิสติค

การวิเคราะห์การถดถอยโลจิสติค แบ่งออกเป็น 2 ประเภท คือ

1.1 Binary Logistic จะใช้เมื่อตัวแปรตาม เป็นตัวแปรเชิงกลุ่มที่มีค่าได้เพียง 2 ค่า เช่น Y=1 ถ้านักศึกษาสอบผ่าน หรือ = 0 ถ้านักศึกษาสอบไม่ผ่าน

1.2 Multinomial Logistic จะใช้เมื่อตัวแปรตาม เป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า เช่น Y = 1 หมายถึงไม่เป็นโรคมะเร็ง Y = 2 หมายถึงการเป็นมะเร็งขั้นต้น ... Y = 5 หมายถึงการเป็นมะเร็งขั้นสุดท้าย

สำหรับในบทนี้จะพิจารณาเฉพาะกรณีที่ตัวแปรตาม Y มีค่าเพียงแค่สองค่า (dichotomous) ซึ่งเรียกว่า Binary Logistic Regression

2. วัตถุประสงค์การวิเคราะห์การถดถอยโลจิสติค

วัตถุประสงค์ของการวิเคราะห์การถดถอยโลจิสติคมาใช้ในการวิเคราะห์ก็เพื่อจะทำนายว่าเหตุการณ์หนึ่งจะเกิดขึ้นได้หรือไม่ หรือจะมีโอกาสเกิดขึ้นมากน้อยเพียงใด โดยกำหนดว่ามีตัวแปรตัวหนึ่งหรือหลายตัวที่ส่งผลต่อการเกิดเหตุการณ์นั้นๆทำให้เราทราบเหตุผลของการเกิดเหตุการณ์นั้นหรือไม่เกิดเหตุการณ์นั้น สำหรับ Binary Logistic Regression ที่จะกล่าวถึงในบทนี้ค่าของตัวแปรตามที่ปรากฏเป็นข้อมูลสำหรับวิเคราะห์จะมีเพียง 2 ค่าเท่านั้น กล่าวคือ ใช่ (Yes) หรือไม่ใช่ (No) ซึ่งต่อไปจะแทนค่าด้วย 0 และ 1 ตามลำดับ

กำหนดให้ odds = P / (1-P)

โดยค่า odds จะใช้แทนอัตราส่วนของความน่าจะเป็นของการเกิดเหตุการณ์และความน่าจะเป็นของการไม่เกิดเหตุการณ์นั้น เช่น ถ้า odds มีค่าเท่ากับ 6 แสดงว่าโอกาสที่จะเกิดเหตุการณ์ที่สนใจเป็น 6 เท่าของโอกาสที่จะไม่เกิดเหตุการณ์

3. เงื่อนไขในการวิเคราะห์การถดถอยโลจิสติค

การวิเคราะห์การถดถอยโลจิสติคที่จะกล่าวถึงในบทนี้เป็น Binary Logistic Regression ดังนั้นตัวแปรตามจึงมีเพียง 2 ค่า (Dichotomous Variable) คือเป็น 0 และ 1 ส่วนตัวแปรอิสระอาจมีเพียงตัวเดียวหรือหลายตัวก็ได้ โดยอาจจะเป็นตัวแปรต่อเนื่อง (Continuous Variable) ตัวแปรจำแนกประเภท (Categorical Variables) หรือ ตัวแปรที่มี 2 ค่า (Dichotomous Variable) ก็ได้ และในการวิเคราะห์การถดถอยเชิงเส้น จะใช้วิธีกำลังสองน้อยที่สุดในการประมาณค่าสัมประสิทธิ์การถดถอย แต่ในการวิเคราะห์การถดถอยโลจิสติคจะใช้วิธีการแม็กซิมั่มไลค์ลิฮูดในการประมาณค่าสัมประสิทธิ์การถดถอย

4. การประมาณตัวแบบการถดถอยโลจิสติค

วิธีการวิเคราะห์ข้อมูลด้วยเทคนิค Logistic Regression เป็นการประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ในกรณีที่มีตัวแปรอิสระเพียงตัวเดียว Logistic Regression Model สามารถเขียนได้เป็น

$$\text{Prob (event)} = \frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}}$$

$$\text{หรือ Prob (event)} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

เมื่อ $\beta 0$ และ $\beta 1$ หมายถึง ค่าพารามิเตอร์ เมื่อทำการประมาณค่าได้

X หมายถึง ตัวแปรอิสระ

e หมายถึง ค่า natural logarithm ในทางคณิตศาสตร์มีค่าประมาณ 2.71828

ทำนองเดียวกัน ถ้าในกรณีที่มีตัวแปรอิสระมากกว่า 1 ตัวแปร ตัวแบบจะสามารถเขียนได้เป็น

$$\text{Prob (event)} = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_p X_p)}}$$

$$\text{หรือ Prob (event)} = \frac{1}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_p X_p)}} = \pi$$

เมื่อได้ค่าความน่าจะเป็นของการเกิดเหตุการณ์ก็จะสามารถคำนวณค่าความน่าจะเป็นของการไม่เกิดเหตุการณ์นั้นได้คือ

$$\text{Prob (no event)} = 1-\frac{1}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_p X_p)}} = 1-\pi$$

และจากตัวแบบเราสามารถแปลงให้เป็นอัตราส่วนของความน่าจะเป็นระหว่างการเกิดเหตุการณ์กับไม่เกิด หรือเรียกว่าค่า Odds Ratio ได้ดังนี้

$$\text{Odds} = \frac{\pi}{1-\pi} = e^{Z} = e^{(\beta_0, \beta_1 X_1, \ldots\ldots \beta_p X_p)} = e^{\beta_0} e^{\beta_1 X_1} \ldots\ldots e^{\beta_p X_p}$$

5. ตัวอย่างและขั้นตอนการใช้โปรแกรมสำเร็จรูปทางสถิติในการวิเคราะห์การถดถอยโลจิสติค

ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลเกี่ยวกับคนไข้หญิงสมรสแล้วที่ไปตรวจที่โรงพยาบาลเกี่ยวกับโรคมะเร็งปากมดลูก สมมุติว่าในการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อต้องทำนายโอกาสของคนไข้ที่จะเป็นโรคมะเร็งปากมดลูก จากอายุของคนไข้ จำนวนบุตร

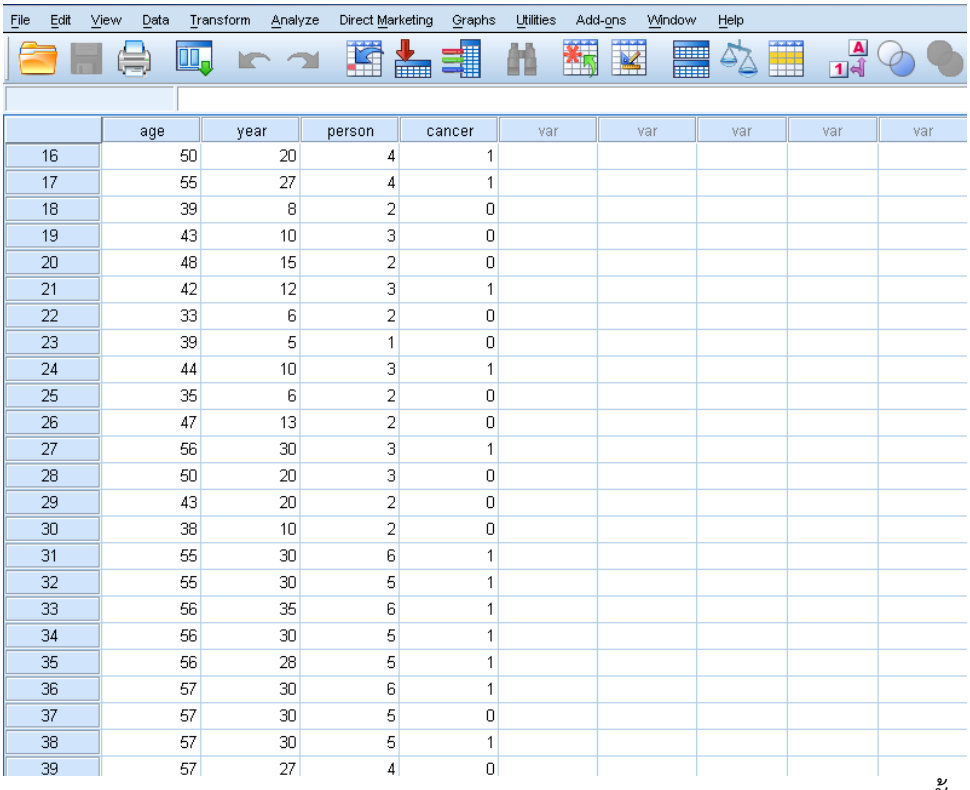
ข้อมูลของคนไข้จำนวน 50 คน ดังนี้

คนไข้คนที่	อายุ	จำนวนปีที่สมรส	จำนวนบุตร	การเป็นโรคมะเร็งปากมดลูก (1 = เป็น , 0 = ไม่เป็น)
1	24	2	1	0
2	33	6	2	0
3	35	7	2	0
4	49	15	3	1
5	30	5	1	0
6	27	3	1	0
7	22	2	1	0
8	50	25	4	1
9	55	30	2	1
10	23	2	1	0
11	44	10	2	1
12	37	10	2	0
13	39	12	3	0
14	32	8	2	0
15	45	15	3	1
16	50	20	4	1
17	55	27	4	1
18	39	8	2	0
19	43	10	3	0
20	48	15	2	0
21	42	12	3	1
22	33	6	2	0

คนไข้คนที่	อายุ	จำนวนปีที่สมรส	จำนวนบุตร	การเป็นโรคมะเร็งปากมดลูก (1 = เป็น , 0 = ไม่เป็น)
23	39	5	1	0
24	44	10	3	1
25	35	6	2	0
26	47	13	2	0
27	56	30	3	1
28	50	20	3	0
29	43	20	2	0
30	38	10	2	0
31	55	30	6	1
32	55	30	5	1
33	56	35	6	1
34	56	30	5	1
35	56	28	5	1
36	57	30	6	1
37	57	30	5	0
38	57	30	5	1
39	57	27	4	0
40	57	28	3	0
41	57	25	3	0
42	58	30	4	0
43	43	22	3	0
44	38	11	2	0
45	47	10	3	0
46	59	15	4	1
47	56	23	2	0
48	50	29	1	0
49	43	20	2	0
50	38	10	2	0

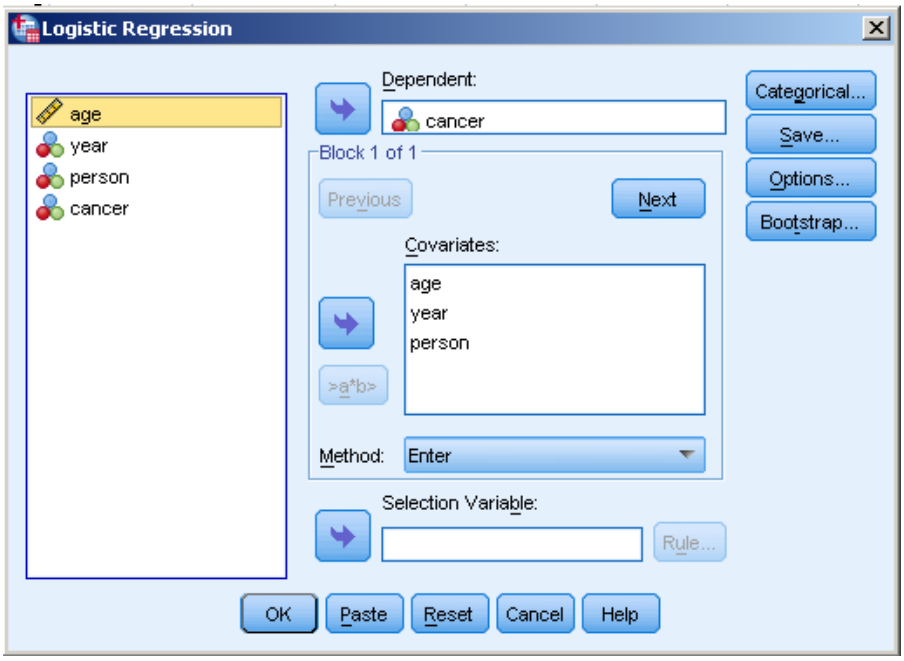
วิเคราะห์ข้อมูลโดยมีขั้นตอน ดังนี้

1. สร้างข้อมูลดังตารางลงในโปรแกรมจากนั้นให้บันทึกข้อมูลโดยใช้ชื่อ file ว่า Binary Logistics ดังภาพ

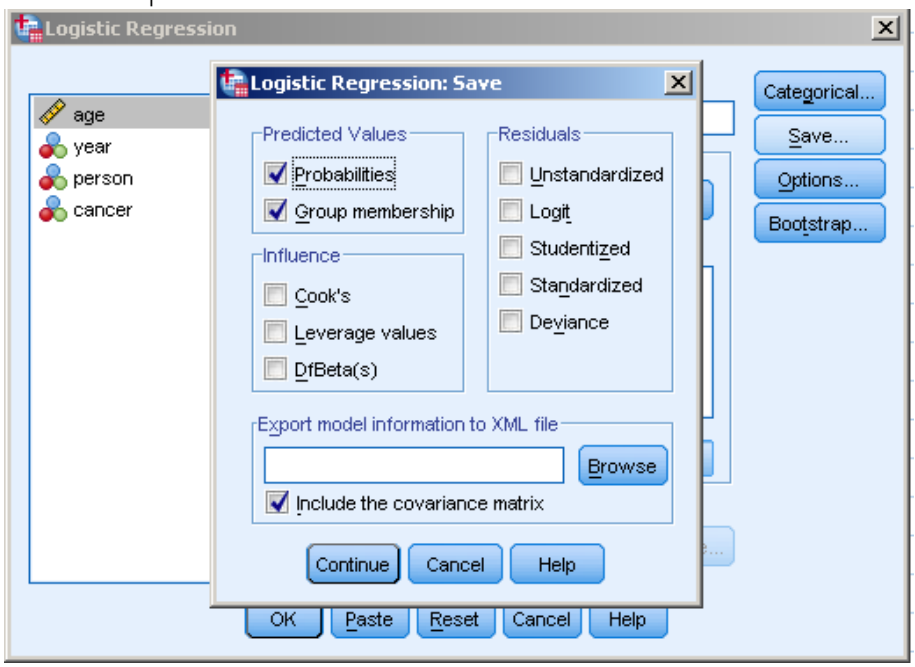


	age	year	person	cancer	var	var	var	var	var
16	50	20	4	1					
17	55	27	4	1					
18	39	8	2	0					
19	43	10	3	0					
20	48	15	2	0					
21	42	12	3	1					
22	33	6	2	0					
23	39	5	1	0					
24	44	10	3	1					
25	35	6	2	0					
26	47	13	2	0					
27	56	30	3	1					
28	50	20	3	0					
29	43	20	2	0					
30	38	10	2	0					
31	55	30	6	1					
32	55	30	5	1					
33	56	35	6	1					
34	56	30	5	1					
35	56	28	5	1					
36	57	30	6	1					
37	57	30	5	0					
38	57	30	5	1					
39	57	27	4	0					

2. Click Analyze ⟶ Regression ⟶ Binary Logistics จากนั้นเลือกตัวแปร cancer ไปไว้ที่ช่อง Dependent และ นำตัวแปร age ,year และ person ใส่ไว้ที่ช่อง Covariate ดังภาพ



3. Click Save ในส่วนของ Predicted Value เลือก Probabilities และ Group membership ดังภาพ



4. Click Continue แล้ว OK จะปรากฏผลลัพธ์ ดังตาราง

ตารางที่ 4.1 ตารางแสดงจำนวนข้อมูลที่นำมาวิเคราะห์

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	50	100.0
	Missing Cases	0	.0
	Total	50	100.0
Unselected Cases		0	.0
Total		50	100.0

a. If weight is in effect, see classification table for the total number of cases.

จากตารางที่ 4.1 บอกเพียงแค่ว่ามีจำนวนข้อมูลเท่าใด และเมื่อคิดเป็นร้อยละจะได้เท่าใด นอกจากนี้ยังบอกอีกว่ามี missing case หรือไม่

ตารางที่ 4.2 ตารางแสดงค่าของตัวแปร

Dependent Variable Encoding

Original Value	Internal Value
ไม่เป็น	0
เป็น	1

จากตาราง 4.2 บอกให้ทราบว่าตัวแปรตามมีค่าเป็นอะไร ในที่นี้มีค่าเป็น 0 และ 1

ตารางที่ 4.3 ตารางแสดงตัวแบบการถดถอย

Block 0: Beginning Block

Classification Table[a,b]

Obse□ved			Predicted		
			cancer		Percentage Correct
			ไม่เป็น	เป็น	
Step 0	cancer	ไม่เป็น	32	0	100.0
		เป็น	18	0	.0
	Overall Percentage				64.0

a. Constant is included in the model.

b. The cut value is .500

จากตารางที่ 4.3 ตารางนี้ไม่ค่อยมีความสำคัญมากนักเพราะว่าตัวแบบการถดถอยที่ใช้สำหรับตารางนี้มีแต่ ค่าคงที่เท่านั้น ไม่มีตัวแปรอิสระอยู่ในตัวแบบเลย (ดูตรงหมายเหตุ a) ดังนั้นเราจะข้ามไปและจะไม่สนใจค่าทั้งหลายที่อยู่ในตาราง สำหรับหมาย b หมายความว่า เมื่อคำนวณได้ค่าความน่าจะ เป็น = .50 หรือมากกว่าก็จะจำแนกให้อยู่กลุ่ม 1 ถ้าได้ค่าน้อยกว่าก็จะจำแนกให้อยู่ในกลุ่ม 0

ตารางที่ 4.4 ตารางแสดงค่าสัมประสิทธิ์การถดถอยของค่าคงที่

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.575	.295	3.814	1	.051	.563

จากตารางที่ 4.4 เป็นตารางแสดงค่าเกี่ยวกับค่าสัมประสิทธิ์การถดถอยของค่าคงที่ อย่างไรก็ตามตารางนี้ไม่มี ความสำคัญเท่าใดนักเช่นกัน เพราะว่าในตัวแบบมีแค่ค่าคงที่เท่านั้น ไม่มีตัวแปรอิสระเลย ดังนั้น เราจะข้ามตารางนี้ไป

ตารางที่ 4.5 ตารางแสดงเพื่อบอกให้ทราบว่าตัวแปรอิสระใดบ้างที่ไม่ได้อยู่ในตัวแบบ

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	age	13.189	1	.000
		year	11.407	1	.001
		person	19.977	1	.000
	Overall Statistics		20.362	3	.000

จากตารางที่ 4.5 บอกให้ทราบว่าตัวแปรอิสระใดบ้างที่ไม่ได้อยู่ในตัวแบบ ซึ่งในตารางนี้บอกว่าตัวแปร age year และ person ไม่อยู่ในตัวแบบ สำหรับ Score เป็นตัวทดสอบสถิติที่ใช้ในการ พยากรณ์ว่าตัวแปรอิสระนั้น ๆจะมีนัยสำคัญเพียงพอหรือไม่ที่อยู่ในตัวแบบ ซึ่งในที่นี้ ก็จะเห็นว่า ตัวแปร age year และ person มีนัยสำคัญทางสถิติ (ค่า p - Value < .05)

ตารางที่ 4.6 ตารางแสดงค่า Omnibus Tests of Model Coefficients บอกผลของการทดสอบตัวแบบการถดถอยโลจิสติค

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.945	3	.000
	Block	23.945	3	.000
	Model	23.945	3	.000

โดยทั่วไปแล้ว output ส่วนนี้เป็นต้นไปเป็นสิ่งที่นักวิจัยสนใจที่จะตีความหมาย Omnibus Tests of Model Coefficients จะบอกให้ทราบผลของการทดสอบตัวแบบการถดถอยโลจิสติค โดยที่

Step 1 เป็นขั้นตอนแรก (หรือตัวแบบแรก) ที่มีตัวแปรอิสระอยู่ในตัวแบบ ในกรณีนี้เรา กำหนดตัวแบบของเราให้เป็นตัวแบบที่เต็มรูปแบบ (full model) ผู้วิจัยอาจจะกำหนดให้มี step มากกว่านี้ก็ได้ถ้าหาก ใช้วิธีการที่เรียกว่า stepwise หรือมีการใช้ blocking of variables ค่า chi-square ในคอลัมน์ที่สองของตารางเป็นค่าสถิติที่ใช้ทดสอบความมีนัยสำคัญของ step, model และ block สำหรับในการวิเคราะห์นี้ค่า

Step 1 เป็นขั้นตอนแรก (หรือตัวแบบแรก) ที่มีตัวแปรอิสระอยู่ในตัวแบบ ในกรณีนี้เรา กำหนดตัวแบบของเราให้เป็นตัวแบบที่เต็มรูปแบบ (full model) ผู้วิจัยอาจจะกำหนดให้มี step มากกว่านี้ก็ได้ ถ้าหาก ใช้วิธีการที่เรียกว่า stepwise หรือมีการใช้ blocking of variables ค่า chi-square ในคอลัมน์ที่สองของตารางเป็นค่าสถิติที่ใช้ทดสอบความมีนัยสำคัญของ step, model และ block สำหรับในการวิเคราะห์นี้ค่า chi-square จะเท่ากันหมด เพราะว่าการ ถดถอยโลจิสติคไม่ได้ใช้แบบ stepwise หรือ block ค่า chi-square สำหรับตัวแบบในตาราง F มี ค่าเท่ากับ 23.945 ค่า P - Value น้อยกว่า .05 แสดงว่า โดยภาพรวมของตัวแบบ สัมประสิทธิ์การถดถอย ตัวแปรอิสระที่มีอยู่ในตัวแบบอย่างน้อยที่สุดหนึ่งตัวไม่เท่ากับ 0

ตารางที่ 4.7 ตารางแสดงค่าสถิติ –2 log likelihood ของตัวแบบสุดท้าย (final model)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	41.397[a]	.381	.522

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

จากตารางที่ 4.7 บอกให้ทราบเกี่ยวกับค่าสถิติ –2 log likelihood ของตัวแบบสุดท้าย (final model) ค่าตัวเลขของเทอมดังกล่าวไม่บอกข่าวสารอะไรมากนัก แต่เราสามารถนำเอาค่าสถิติ –2 log likelihood ไปใช้ในการเปรียบเทียบระหว่างตัวแบบที่ลดรูป (reduced model)ได้ ค่าสถิติในคอลัมน์ 2 และ 3 เป็น Pseudo R-square จริง ๆแล้วในการถดถอยโลจิสติคจะไม่มีค่า R-square เหมือนค่า R-square ในการถดถอยเชิงเส้น ดังนั้นการแปลความหมายของ Pseudo R-square ในคอลัมน์ 2 และ 3 ควรจะต้องแปลด้วยความระมัดระวังหรือถ้าหลีกเลี่ยงการแปลได้ก็ควรหลีกเลี่ยง

ตารางที่ 4.8 ตารางแสดงการจำแนกกลุ่มสำหรับค่าที่สังเกตได้กับค่าที่พยากรณ์

Classification Table[a]

Observed			Predicted		
			cancer		Percentage Correct
			ไม่เป็น	เป็น	
Step 1	cancer	ไม่เป็น	29	3	90.6
		เป็น	7	11	61.1
	Overall Percentage				80.0

a. The cut value is .500

จากตารางที่ 4.8 เป็นการตารางแสดงการจำแนกกลุ่มสำหรับค่าที่สังเกตได้กับค่าที่พยากรณ์ในตารางบอกให้ทราบว่า

1. ค่าที่สังเกตได้ว่า Cancer = 0 ตัวแบบพยากรณ์ว่า Cancer = 0 มีจำนวน = 29 คำนวณได้ถูกต้องร้อยละ 90.6
2. ค่าที่สังเกตได้ว่า Cancer = 0 ตัวแบบพยากรณ์ว่า Cancer = 1 มีจำนวน = 3
3. ค่าที่สังเกตได้ว่า Cancer = 1 ตัวแบบพยากรณ์ว่า Cancer = 0 มีจำนวน = 7
4. ค่าที่สังเกตได้ว่า Cancer = 1 ตัวแบบพยากรณ์ว่า Cancer = 1 มีจำนวน = 11 คำนวณได้ถูกต้องร้อยละ 61.1

โดยรวมแล้วตัวแบบการถดถอยที่มี age ,year และ person เป็นตัวแปรอิสระจำแนกคนไข้ได้ถูกต้องตามกลุ่ม (เป็น โรคมะเร็งปากมดลูกกับไม่เป็น) ร้อยละ 80.0

ตารางที่ 4.9 ตารางแสดงค่าสัมประสิทธิ์การถดถอยกับตัวสถิติที่ใช้ทดสอบ

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	.104	.104	1.007	1	.316	1.110
	year	-.077	.099	.605	1	.437	.926
	person	1.201	.507	5.618	1	.018	3.325
	Constant	-7.776	3.541	4.822	1	.028	.000

a. Variable(s) entered on step 1: age, year, person.

จากตารางที่ 4.9 บอกให้ทราบเกี่ยวกับสัมประสิทธิ์การถดถอยกับตัวสถิติที่ใช้ทดสอบและบอกให้ทราบถึงความมีนัยสำคัญทางสถิติของตัวแปรอิสระแต่ละตัว จากตารางจะได้ว่ามีเพียงตัวแปร person (จำนวนบุตร) เท่านั้นที่มีค่า p - Value < .05 แสดงว่า ตัวแปร person (จำนวนบุตร) มีความสำคัญเพียงพอที่จะอยู่ในตัวแบบ โดยจะมีค่าสัมประสิทธิ์การถดถอยเท่ากับ 1.201 สำหรับค่า Exp(B) ในคอลัมน์ขวาสุดคือค่า odds ratio สำหรับ ตัวแปรอิสระ ในที่นี้ odds ratio ของตัวแปร person (จำนวนบุตร) คือ 3.325 ค่านี้มีความหมายว่า ถ้าตัวแปร person (จำนวนบุตร) เพิ่มขึ้นหนึ่งหน่วยก็จะทำให้การเป็นโรคมะเร็งปากมดลูกมีโอกาสมากกว่าการไม่เป็นโรคมะเร็งปากมดลูก 3.325 เท่าข้อมูลจากตารางที่ 4.9 เมื่อนำมาแทนค่าในตัวแบบการถดถอยโลจิสติค จะได้ดังนี้

$$\text{logit}(Y) = -7.776 + 1.201(\text{person}) \quad \text{.....(1)}$$

$$\text{odds}(Y=1) = e^{-7.776 + 1.201(\text{person})}$$

$$= e^{-7.776} e^{1.201(\text{person})} \quad \text{.....(2)}$$

$$P(Y=1) = \frac{e^{-7.776+1.201(\text{person})}}{1+e^{-7.776+1.201(\text{person})}} \quad \text{.....(3)}$$

การอธิบายตัวแบบในสมการ (1) จะเข้าใจยากกว่าการอธิบายโดย ใช้ตัวแบบในสมการ (2) หรือ (3) เพราะสมการ (1) จะต้องอธิบายในรูปของ log odds แต่ถึงแม้ว่าจะเข้าใจยาก เราก็ยังวิเคราะห์ค่าสัมประสิทธิ์ได้ เนื่องจากค่าสัมประสิทธ์การถดถอย = 1.201 บอกได้ว่าเมื่อจำนวนบุตรเพิ่มขึ้นโอกาสการเป็นโรคมะเร็งปากมดลูกก็เพิ่มขึ้นด้วย (เพราะค่าสัมประสิทธิ์การ ถดถอยเป็นบวก)
ในที่นี้จะอธิบายโดยใช้ตัวแบบใน (2) และ (3) จากสมการ (2) บอกให้ทราบว่า สำหรับทุก ๆ X ที่เพิ่มขึ้นหนึ่งหน่วย ค่า odds เพิ่มขึ้นเป็นพหุคูณของ e^{β_1} นั่นคือ สำหรับจำนวนบุตรที่เพิ่มขึ้นทุก ๆ หนึ่งคน โอกาสการเป็นโรคมะเร็งปากมดลูกจะเพิ่มขึ้นเป็นพหุคูณของ $e^{1.201}$ (= 3.323) หรือกล่าวอีกนัยหนึ่งได้ว่า โอกาส(หรือความเสี่ยง)ของการเป็นโรคมะเร็งปากมดลูก ณ ที่จำนวนบุตร x+1 คน เท่ากับ โอกาส(หรือความเสี่ยง) ของการเป็นโรคมะเร็งปากมดลูก ณ ที่จำนวนบุตร x ปี คูณด้วย . $e^{1.201}$ ($e^{1.201}$ มีค่าเท่ากับ 3.323)
<u>เช่น</u> ถ้า จำนวนบุตร = 6 คน odds(Y=1) มีค่าเท่ากับ $e^{-7.776}e^{1.201(6)}$ = (0.00041)(3.332)^6= 0.561 นั่นคือ คนที่มีบุตร 6 คน มีความเสี่ยงของการเป็นโรคมะเร็งปากมดลูก มากกว่าคนที่มีบุตร 5 คน 0.561 เท่า

ถ้าหากจะอธิบายด้วยสมการที่ (3) จะได้ว่า เมื่อคนเรามีมีบุตร 5 คน ความน่าจะเป็นที่จะเป็นโรคมะเร็งปากมดลูก = $\frac{e^{-7.776+1.201(5)}}{1+e^{-7.776+1.201(5)}} = 0.145$ และเมื่อคนเรามีบุตร 6 คน ความน่าจะเป็นที่จะเป็นโรคหัวใจ = $\frac{e^{-7.776+1.201(6)}}{1+e^{-7.776+1.201(6)}} = 0.360$ (แสดงว่าจำนวนบุตรมากขึ้นความน่าจะเป็นที่จะเป็นโรคมะเร็งปากมดลูกก็จะสูงขึ้น)

6. บทสรุป

ในปัจจุบันการใช้เทคนิคทางสถิติขั้นสูงในการวิเคราะห์ข้อมูลเพื่อการวิจัยเป็นที่ยอมรับว่าให้ผลสรุปที่มีความน่าเชื่อถือมากยิ่งขึ้น Logistic Regression Analysis เป็นทางเลือกหนึ่งที่สามารถทำให้ผู้วิจัยสามารถบรรลุวัตถุประสงค์ที่ตั้งไว้ได้ ถ้าข้อมูลตัวแปรตามที่ใช้มีลักษณะเป็นตัวแปรเชิงกลุ่ม (Dichotomous) ซึ่งบ่อยครั้งที่จะพบว่าในสถานการณ์ต่างๆข้อมูลที่นำมาวิเคราะห์มักจะมีตัวแปรตามเป็นตัวแปรเชิงกลุ่ม โดยในการวิเคราะห์การถดถอยโลจิสติคจะใช้วิธีการแม็กซิมั่มไลค์ลิฮูดในการประมาณค่าสัมประสิทธิ์การถดถอย และวัตถุประสงค์ของการวิเคราะห์การถดถอยโลจิสติคก็เพื่อจะทำนายว่าเหตุการณ์หนึ่งจะเกิดขึ้นได้หรือไม่ หรือจะมีโอกาสเกิดขึ้นมากน้อยเพียงใด ซึ่งแตกต่างจากการวิเคราะห์ถดถอยทั่วไปที่ใช้วิธีกำลังสองน้อยที่สุดในการประมาณค่าสัมประสิทธิ์การถดถอย และวัตถุประสงค์ของการวิเคราะห์การถดถอยทั่วไปก็คือการประมาณค่าของตัวแปรตาม

แบบฝึกหัดบทที่ 4
เรื่อง การวิเคราะห์การถดถอยโลจิสติค

---

1. จากข้อมูลเกี่ยวกับคนไข้ที่เข้ารับการผ่าตัดที่โรงพยาบาลแห่งหนึ่ง ในการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อต้องการทำนายโอกาสของคนไข้ที่จะเป็นไข้หลังจากเข้ารับการผ่าตัดจากอายุของคนไข้
ข้อมูลของคนไข้จำนวน 50 คน ดังนี้

คนที่	อายุ	การเป็นไข้ (1 = เป็น , 0 = ไม่เป็น)	คนที่	อายุ	การเป็นไข้ (1 = เป็น , 0 = ไม่เป็น)
1	55	1	26	32	1
2	55	1	27	23	0
3	56	0	28	24	1
4	56	0	29	25	0
5	56	0	30	25	0
6	57	0	31	26	0
7	57	0	32	26	0
8	57	1	33	28	1
9	57	1	34	28	1
10	57	1	35	29	1
11	57	1	36	30	0
12	58	0	37	30	0
13	43	0	38	48	0
14	30	0	39	60	1
15	30	0	40	64	1
16	30	0	41	58	1
17	30	1	42	62	0
18	30	0	43	53	0
19	32	1	44	47	0
20	32	0	45	26	1
21	33	0	46	28	1
22	33	0	47	39	0
23	45	1	48	56	0
24	39	0	49	37	1
25	56	1	50	44	1

จงหาตัวแบบการถดถอย Logistics Regression โดยใช้โปรแกรม SPSS พร้อมทั้งอธิบายตัวแบบที่ได้มาโดยละเอียด

2. จากข้อมูลเกี่ยวกับการพ้นสภาพนักศึกษาของนักศึกษาคณะวิทยาศาสตร์มหาวิทยาลัยแห่งหนึ่ง โดยในการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อต้องการทำนายโอกาสการพ้นสภาพนักศึกษาจากตัวแปร เกรดเฉลี่ยสะสมมัธยมศึกษาตอนปลาย จำนวนชั่วโมงในการอ่านหนังสือเรียนต่อสัปดาห์ จำนวนชั่วโมงในการทำงานระหว่างเรียนต่อสัปดาห์ สุ่มนักศึกษาที่ทำงานพิเศษระหว่างเรียนมา 50 คนได้ข้อมูลดังนี้

คนที่	เกรดเฉลี่ยสะสมมัธยมศึกษาตอนปลาย	จำนวนชั่วโมงในการอ่านหนังสือเรียนต่อสัปดาห์	จำนวนชั่วโมงในการทำงานระหว่างเรียนต่อสัปดาห์	การพ้นสภาพนักศึกษา (1 = ไม่พ้นสภาพ , 0 = พ้นสภาพ)
1	3.05	10	15	1
2	2.75	12	15	1
3	2.01	3	20	0
4	2.00	12	20	1
5	1.96	6	15	0
6	3.55	8	12	1
7	1.87	7	12	0
8	3.00	9	15	1
9	2.47	9	20	1
10	2.38	1	20	1
11	1.77	12	15	0
12	3.87	15	12	1
13	2.43	3	15	1
14	1.98	7	20	0
15	3.35	8	15	1
16	3.45	15	12	1
17	2.78	10	20	1
18	1.94	6	15	1
19	2.77	5	12	1
20	3.86	5	10	0
21	2.54	8	10	1
22	3.44	8	12	1
23	2.01	7	15	0
24	3.54	6	10	1

คนที่	เกรดเฉลี่ยสะสมมัธยมศึกษาตอนปลาย	จำนวนชั่วโมงในการอ่านหนังสือเรียนต่อสัปดาห์	จำนวนชั่วโมงในการทำงานระหว่างเรียนต่อสัปดาห์	การพ้นสภาพนักศึกษา (1 = ไม่พ้นสภาพ , 0 = พ้นสภาพ)
25	2.25	7	10	0
26	1.85	8	12	0
27	2.26	10	15	1
28	3.49	10	20	1
29	3.32	12	15	1
30	2.78	12	12	1
31	2.44	9	20	1
32	3.18	5	10	1
33	2.97	6	10	1
34	1.78	7	12	1
35	2.55	7	15	0
36	3.49	7	10	1
37	2.78	8	15	1
38	1.65	15	15	0
39	3.33	10	20	1
40	3.45	3	20	1
41	2.56	4	15	1
42	1.99	4	12	1
43	2.53	5	12	0
44	3.49	8	12	1
45	2.78	6	12	0
46	1.98	15	15	1
47	2.02	15	20	1
48	2.07	14	20	1
49	2.44	13	15	1
50	2.56	10	12	1

จงหาตัวแบบการถดถอย Logistics Regression โดยใช้โปรแกรม SPSS พร้อมทั้งอธิบายตัวแบบที่ได้มาโดยละเอียด

เอกสารอ้างอิงบทที่ 4

กัลยา วานิชย์บัญชา. (2551). **การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows**. พิมพ์ครั้งที่ 6. กรุงเทพฯ: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

บุญชม ศรีสะอาด. (2543). **การวิจัยทางการวัดผลและประเมินผล.** กรุงเทพฯ : สุวีริยาสาส์น.

ศิริชัย พงษ์วิชัย. (2544). **การวิเคราะห์ข้อมูลทางสถิติด้วยคอมพิวเตอร์.** พิมพ์ครั้งที่ 11. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย.

สุชาติ ประสิทธิ์รัฐสินธุ์. 2540. **เทคนิคการวิเคราะห์ตัวแปรหลายตัวสำหรับการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์.** พิมพ์ครั้งที่ 4. กรุงเทพฯ : เลี่ยงเชียง.

Hosmer, D. and Lemeshow, S. 1989. **Applied logistic regression**. John Wiley and Sons, Inc.

การวิเคราะห์การถดถอยโลจิสติค. (2557). ค้นเมื่อ 24 เมษายน 2557, จาก http:// netra.lpru.ac.th/~phaitoon.

การจำแนกกลุ่มตัวแปรด้วยเทคนิค Cluster Analysis. (2557). ค้นเมื่อ 22 เมษายน 2557, จาก http://www.saruthipong.com/port/document/299-705/299-705-10.pdf.