

บทที่ 3

วิธีดำเนินการศึกษา

1) ข้อมูลที่ใช้ในการศึกษาได้จากการ generate ตัวแปรที่ใช้ทำนาย(ตัวแปรอิสระ) ให้มีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ โดยตัวแปรเชิงปริมาณ ได้จากการ generate ให้ตัวแปรมีการแจกแจงปกติมาตรฐานจำนวน n_1 ตัว ส่วนตัวแปรเชิงคุณภาพ ได้จากการ generate ให้ตัวแปรมีการแจกแจงทวินามโดยมีค่าของตัวแปรเท่ากับ 0 กับ 1 โดย $P[X_i = 0] = P[X_i = 1] = 0.5$ กำหนดให้ตัวแปรที่ใช้ทำนายซึ่งเป็นตัวแปรเชิงปริมาณมีความสัมพันธ์กันตั้งแต่ระดับน้อยถึงระดับมาก โดยครอบคลุมประเด็นต่างๆดังนี้

1.1) จำนวนตัวแปรที่ใช้ทำนายแบ่งเป็น 3 ระดับ คือ 4 ตัว 6 ตัว และ 10 ตัว

1.2) กำหนดให้ตัวแปรอิสระมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพโดยกำหนดให้มีตัวแปรเชิงคุณภาพ 3 ระดับคือ (1)ไม่มีตัวแปรเชิงคุณภาพ (2) มีจำนวนตัวแปรเชิงคุณภาพไม่ถึงร้อยละ 50 (3) มีตัวแปรเชิงคุณภาพตั้งแต่ร้อยละ 50 ขึ้นไป

1.3) ความสัมพันธ์ระหว่างตัวแปรที่ใช้ทำนาย ซึ่งวัดด้วยค่า Variance Inflation Factor (VIF) ที่มีค่ามากที่สุด โดยกำหนดให้มีค่า 3 ระดับดังนี้ (1) $VIF \leq 4$ (2) $4 < VIF \leq 10$ (3) $VIF > 10$

จำนวนตัวแปรที่ใช้ในการศึกษา จำนวนตัวแปรเชิงปริมาณ จำนวนตัวแปรเชิงคุณภาพ และค่า VIF ของกรณีศึกษาที่มีทั้งสิ้น 36 กรณี (ตาราง 1.2)

2) generate ตัวแปรตอบสนอง Y จากตัวแปรสุ่มที่มีการแจกแจงทวินามซึ่งมีค่า $P[X_i = 1] = \pi$ โดย $\pi(X) = \frac{\exp[g(X)]}{1 + \exp[g(X)]}$ และ $\beta_1, \beta_2, \dots, \beta_k$ เป็นค่าสัมประสิทธิ์การถดถอย โดย $g(X) = \ln \frac{\pi(X)}{1 - \pi(X)} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$ และกำหนดให้ $\beta_0 = \beta_1 = \beta_2 = \dots = \beta_k = 1$ ซึ่งอธิบายได้ว่าตัวแปรที่ใช้ทำนายแต่ละตัวมีผลต่อตัวแปรตอบสนองไม่แตกต่างกัน

3) สุ่มตัวอย่างจากประชากรโดยให้มียอดส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรที่ใช้ทำนายเท่ากับ 50 ต่อ 1 30 ต่อ 1 20 ต่อ 1 และ 15 ต่อ 1 เพื่อหาตัวแบบในการทำนายการเป็นสมาชิกกลุ่ม

4) แบ่งข้อมูลออกเป็น 2 ชุด โดยข้อมูลชุดที่ 1 ใช้ในการสร้างสมการสำหรับใช้ในการพยากรณ์ ส่วนข้อมูลชุดที่ 2 ใช้ในการหาค่าความคลาดเคลื่อนในการจำแนกกลุ่ม

5) ใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธีวิเคราะห์การถดถอยโลจิสติก วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง และวิธีดริฟต์เรกชัน (โดย วิธี Ridge1 กำหนดค่า k เริ่มต้นด้วยวิธี Hoerl และ Kennard ส่วน วิธี Ridge2 กำหนดค่า k เริ่มต้นด้วยวิธี Lawless และ Wang) ประมาณค่าความน่าจะเป็นที่ค่าสังเกตแต่ละค่าจะถูกจำแนกอยู่ในกลุ่มใดกลุ่มหนึ่งซึ่งหากค่าความ

น่าจะเป็นที่ค่าสังเกตจะถูกจำแนกเป็นกลุ่ม 1 มีค่ามากกว่าค่าความน่าจะเป็นที่ค่าสังเกตจะถูกจำแนกเป็นกลุ่ม 0 ค่าสังเกตค่านั้นจะถูกจำแนกให้อยู่กลุ่ม 1

6) ในการเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ต้องด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีวิธีจรีเกรสชัน วิธีวิเคราะห์การถดถอยโลจิสติก และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงโดยหาค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของตัวบ่งชี้ B ตัวบ่งชี้ C และความคลาดเคลื่อนในการจำแนกกลุ่ม จากการทำซ้ำ 1000 รอบ และ 5000 รอบ (ตาราง 3.1)

ตาราง 3.1 ค่าเฉลี่ย (mean) และค่าส่วนเบี่ยงเบนมาตรฐาน (standard deviation:sd) ของ B C และค่าความคลาดเคลื่อนในการจำแนก (Classification Error:CE) ที่ได้จากวิธีวิเคราะห์การถดถอยพหุคูณ (B.OLS , C.OLS , CE.OLS) วิธีวิธีจรีเกรสชันซึ่งกำหนดค่า k เริ่มต้นด้วยวิธีที่ 1 และวิธีที่ 2 (B.Ridge1,C.Ridge1, CE.Ridge1 และ B.Ridge2, C.Ridge2, CE.Ridge2) วิธีวิเคราะห์การถดถอยโลจิสติก (B.Logit , C.Logit, CE.Logit) และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง(B.LDA, C.LDA, CE.LDA) จากการทำซ้ำ 1000 รอบ และ 5000 รอบ ครั้งที่ 1 และ ครั้งที่ 2 เมื่อกำหนดให้มีจำนวนตัวแปรเชิงปริมาณ 4 ตัว ไม่มีตัวแปรเชิงคุณภาพ

ตัวบ่งชี้	จำนวนซ้ำ=1000 ครั้งที่ 1		จำนวนซ้ำ=1000 ครั้งที่ 2		จำนวนซ้ำ=5000 ครั้งที่ 1		จำนวนซ้ำ=5000 ครั้งที่ 2	
	mean	sd	mean	sd	mean	sd	mean	sd
B.OLS	0.84140	0.01233	0.84183	0.01275	0.84191	0.01255	0.84214	0.01227
B.Ridge1	0.84109	0.01219	0.84151	0.01262	0.84159	0.01242	0.84181	0.01214
B.Ridge2	0.84109	0.01219	0.84151	0.01261	0.84159	0.01242	0.84181	0.01214
B.Logit	0.84338	0.01531	0.84385	0.01575	0.84405	0.01549	0.84431	0.01509
B.LDA	0.84352	0.01478	0.84405	0.01521	0.84423	0.01500	0.84447	0.01462
C.OLS	0.85524	0.02604	0.85627	0.02685	0.85651	0.02637	0.85703	0.02588
C.Ridge1	0.85526	0.02604	0.85629	0.02684	0.85652	0.02636	0.85704	0.02588
C.Ridge2	0.85526	0.02604	0.85629	0.02684	0.85652	0.02636	0.85704	0.02588
C.Logit	0.85560	0.02592	0.85650	0.02684	0.85677	0.02632	0.85733	0.02588
C.LDA	0.85533	0.02601	0.85637	0.02683	0.85660	0.02635	0.85712	0.02587
CE.OLS	0.22956	0.03004	0.22876	0.02968	0.22877	0.03056	0.22868	0.02949
CE.Ridge1	0.22968	0.03010	0.22874	0.02969	0.22887	0.03063	0.22877	0.02946
CE.Ridge2	0.22970	0.03008	0.22876	0.02966	0.22886	0.03062	0.22877	0.02945
CE.Logit	0.22917	0.03056	0.22881	0.02976	0.22862	0.03044	0.22847	0.02956
CE.LDA	0.22956	0.03001	0.22879	0.02969	0.22878	0.03057	0.22868	0.02946

7) คำนวณค่า P-value ของการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของตัวบ่งชี้ B ตัวบ่งชี้ C และค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มโดย (ตาราง 3.3) จากการใช้ระดับนัยสำคัญ 0.05พบว่าไม่มีความแตกต่างระหว่างค่าเฉลี่ยจากการทำซ้ำ 1000 รอบ และ 5000 รอบ ในครั้งที่ 1 และครั้งที่ 2

ตาราง 3.2 ค่า P-value จากการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ B C และค่าความคลาดเคลื่อนในการจำแนกจากการใช้วิธีวิเคราะห์การถดถอยพหุคูณ (B.OLS , C.OLS , CE.OLS) วิธีรีดจ์เรสชันซึ่งกำหนดค่า k เริ่มต้นด้วยวิธีที่ 1 และวิธีที่ 2 (B.Ridge1,C.Ridge1, CE.Ridge1 และ B.Ridge2,C.Ridge2, CE.Ridge2) วิธีวิเคราะห์การถดถอยโลจิสติก (B.Logit , C.Logit, CE.Logit) และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง (B.LDA, C.LDA, CE.LDA) จากการทำซ้ำ 1000 รอบ และ 5000 รอบ ครั้งที่ 1 และ ครั้งที่ 2 เมื่อกำหนดให้มีจำนวนตัวแปรเชิงปริมาณ 4 ตัว ไม่มีตัวแปรเชิงคุณภาพ

ตัวบ่งชี้	P-value			
	ทดสอบความแตกต่างระหว่างค่าเฉลี่ยที่ได้จาก 1000 รอบในครั้งที่ 1 และครั้งที่ 2	ทดสอบความแตกต่างระหว่างค่าเฉลี่ยที่ได้จาก 5000 รอบในครั้งที่ 1 และครั้งที่ 2	ทดสอบความแตกต่างระหว่างค่าเฉลี่ยที่ได้จาก 1000 รอบในครั้งที่ 1 และ 5000 รอบในครั้งที่ 1	ทดสอบความแตกต่างระหว่างค่าเฉลี่ยที่ได้จาก 1000 รอบในครั้งที่ 2 และ 5000 รอบในครั้งที่ 2
B.OLS	0.44646	0.36052	0.23557	0.48274
B.Ridge1	0.44698	0.36865	0.24153	0.49681
B.Ridge2	0.44474	0.36946	0.23955	0.49744
B.Logit	0.50037	0.39251	0.20690	0.39164
B.LDA	0.43497	0.42238	0.16998	0.42421
C.OLS	0.38065	0.31514	0.16031	0.41298
C.Ridge1	0.38355	0.31946	0.16120	0.41353
C.Ridge2	0.38335	0.31930	0.16145	0.41410
C.Logit	0.44242	0.27798	0.19425	0.37012
C.LDA	0.37736	0.32179	0.15822	0.41797
CE.OLS	0.54918	0.87958	0.45135	0.94024
CE.Ridge1	0.48206	0.86526	0.43746	0.97981
CE.Ridge2	0.48173	0.87831	0.42545	0.98834
CE.Logit	0.78958	0.80778	0.60320	0.74419
CE.LDA	0.56415	0.86119	0.45513	0.91163

8) ทดสอบความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกค่าเฉลี่ยของ B และค่าเฉลี่ยของ C ซึ่งได้จากการใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธีรีดจ์เรสชัน วิธีวิเคราะห์การถดถอยโลจิสติก และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงโดยทำซ้ำ 1000 ครั้ง จากข้อมูลที่ได้จำลองขึ้นในตาราง 3.1

9) สรุปผลการวิจัย