บทที่ 2

ระเบียบวิธีวิจัย

การวิจัยเรื่องนี้มุ่งที่จะศึกษาหาสาเหตุหรือปัจจัยที่มีผลต่อสัมฤทธิ์ผลทางการศึกษาในระดับอุดมศึกษาของสถาบันอุดมศึกษาเอกชน โดยเก็บรวบรวมข้อมูลจากนักศึกษาที่เข้าศึกษาตั้งแต่ปี 2523 ของวิทยาลัยเอกชน 7 แห่ง คือ วิทยาลัยกรุงเทพ วิทยาลัยหอการค้า วิทยาลัยธุรกิจบัณฑิต วิทยาลัยเทคนิคสยาม วิทยาลัยเอเชียอาคเนย์ วิทยาลัยเกริก วิทยาลัยอัสสัมชัญบริหารธุรกิจ ซึ่งข้อมูลที่ได้นี้แบ่งออกเป็น 2 ส่วนคือ ข้อมูลเกี่ยวกับสถานภาพส่วนตัว ซึ่งได้แก่ เพศ สาขาวิชาที่เรียน สาขาที่สำเร็จในระดับมัธยมศึกษาตอนปลาย สาขาที่สำเร็จในระดับอุดมศึกษา เกรดเฉลี่ยสะสมที่ได้จากมัธยมศึกษาตอนปลาย เกรดเฉลี่ยสะสมที่ได้ในระดับอุมศึกษา อีกส่วนหนึ่งของข้อมูลคือ ข้อมูลที่เกี่ยวกับปัญหาส่วนตัวหรือสภาพแวดล้อมของนักศึกษาซึ่งได้แก่ การแบ่งเวลาในการเรียน ความสนใจในการเรียน ทัศนคติต่อวิชาและคณะที่เรียน ความรู้พื้นฐานและสติปัญญา การร่วมกิจกรรมนักศึกษา การปรับตัวในการเรียนด้านเศรษฐกิจ ด้านครอบครัว ด้านสุขภาพและทัศนคติเกี่ยวกับครูผู้สอน

2.1 ประชากรที่ศึกษา

ประชากรที่ใช้ในการศึกษาคือ นักศึกษาที่เข้าศึกษาในวิทยาลัยเอกชนทั้ง 7 แห่งดังกล่าว ที่เข้าศึกษาในปี พ.ศ. 2523 ซึ่งแบ่งเป็น 2 กลุ่มคือ กลุ่มที่สำเร็จตามหลักสูตรและกลุ่มที่ไม่สำเร็จการศึกษารายละเอียดเกี่ยวกับจำนวนนักศึกษาทั้งหมดจำแนกตามสาขาวิชาและสถาบันการศึกษาแสดงไว้ในตารางที่ 1

ตารางที่ 1 จำนวนนักศึกษาจำแนกตามสาขาวิชาการศึกษาและสถาบันการศึกษาได้ดังนี้

วิทยาลัยที่ศึกษา / สาขาวิชาที่ศึกษา	วิทยาลัยกรุงเทพ		วิทยาลัยหอการค้าไทย		วิทยาลัยธุรกิจบัณฑิต		วิทยาลัยเกริก		วิทยาลัยอัสสัมชัญ		วิทยาลัยเอเชียอาคเนย์		วิทยาลัยเทคนิคสยาม		รวมทุกวิทยาลัย	
	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ
สาขาบริหารธุรกิจ	217	173	360	304	166	151	134	68	353	172	148	53	156	58	1534	979
สาขาบัญชี	82	162	152	174	126	94	148	43	360	146	86	42	32	27	987	688
สาขานิเทศศาสตร์	134	168	-	-	-	-	-	-	-	-	-	-	-	-	134	168
รวมทุกสาขาวิชา	433	503	513	478	292	245	282	111	713	318	234	95	188	85	2655	1835

2.2 การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลมาใช้ในการวิจัยครั้งนี้ใช้วิธีการสำรวจจากตัวอย่าง ซึ่งมีขั้นตอนที่สำคัญในการรวบรวมข้อมูลดังนี้

1. จากจำนวนนักศึกษา 4,490 คน ซึ่งเข้าศึกษาในปีการศึกษา 2523 ของสถาบันการศึกษาเอกชนทั้ง 7 แห่ง ซึ่งประกอบด้วยนักศึกษาที่สำเร็จการศึกษา 2655 คน และไม่สำเร็จการศึกษา 1835 คน ผู้วิจัยได้คำนวณขนาดตัวอย่างที่จะต้องใช้ในการวิจัยโดยใช้สูตรดังนี้

$$n = \frac{z_{\frac{\alpha}{2}}^{2} \; s^{2}}{d}$$

เมื่อ $z_{\frac{\alpha}{2}}$ คือ ค่าจากตารางการแจกแจงแบบปกติมาตรฐานที่ระดับนัยสำคัญ α

s คือ ส่วนเบี่ยงเบนมาตรฐานของสัมฤทธิ์ผลทางการศึกษาเมื่อสำเร็จการศึกษาหรือพ้นสภาพนักศึกษา

d คือ ความผิดพลาดของสัมฤทธิ์ผลทางการศึกษาที่ยอมให้เกิดขึ้นได้

เนื่องจากการวิจัยครั้งนี้มีตัวแปรทั้งหมด 62 ตัว ดังนั้นขนาดตัวอย่างที่ใช้จึงอาจมีได้ทั้งหมดถึง 62 ค่า ซึ่งแตกต่างกัน แต่ผู้วิจัยจะใช้ขนาดตัวอย่างที่คำนวณจากตัวแปรที่มีอิทธิพลต่อสัมฤทธิ์ผลทางการศึกษามากที่สุด จากการศึกษาว่าตัวแปรใดที่ควรมีอิทธิพลกับสัมฤทธิ์ผลทางการศึกษาในระดับอุดมศึกษามากที่สุด ปรากฏว่าตัวแปรที่น่าจะมีอิทธิพลกับสัมฤทธิ์ผลทางการศึกษาในระดับอุดมศึกษามากที่สุดคือเกรดเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลาย ผู้วิจัยยังศึกษาต่ออีกว่าความแปรปรวนของเกรดเฉลี่ยสะสมเมื่อแยกตามสาขาและแยกตามวิทยาลัยของนักศึกษาที่สำเร็จการศึกษากับนักศึกษาที่ไม่สำเร็จการศึกษาจะมีความแปรปรวนมากกว่าหรือน้อยกว่าความแปรปรวนของเกรดเฉลี่ยสะสมที่รวมทุกสาขาและทุกวิทยาลัยของนักศึกษาที่สำเร็จการศึกษากับนักศึกษาที่ไม่สำเร็จการศึกษาปรากฏว่า ความแปรปรวนของเกรดเฉลี่ยสะสมเมื่อคิดรวมทุกสาขาและทุกวิทยาลัยมีมากกว่าความแปรปรวนของเกรดเฉลี่ยสะสมเมื่อแยกตามสาขาและแยกตามวิทยาลัย ดังนั้นผู้วิจัยจึงคำนวณหาขนาดตัวอย่างของนักศึกษาที่สำเร็จการศึกษาและไม่สำเร็จการศึกษาโดยใช้ความแปรปรวนของเกรดเฉลี่ยสะสมเมื่อรวมทุกสาขาและทุกวิทยาลัย ซึ่งเท่ากับ 0.549 และ 0.507 ตามลำดับ จากสูตรการหาขนาดตัวอย่างข้างต้น ถ้ากำหนดให้ระดับนัยสำคัญเท่ากับ 0.05 และความผิดพลาดของสัมฤทธิ์ผลทางการศึกษาไม่เกินร้อยละ 5 จะได้ขนาดตัวอย่างของนักศึกษาที่

สำเร็จการศึกษา

$$n = \left(\frac{1.96 \times 0.549}{0.05}\right)^2 = 463 \text{ คน}$$

หรือคิดเป็นร้อยละ $\frac{463 \times 100}{2655} = 17$ ของนักศึกษาซึ่งเข้าศึกษาในปีการศึกษา 2523 ที่สำเร็จการศึกษา

สำหรับขนาดตัวอย่างของนักศึกษาที่ไม่สำเร็จการศึกษาสามารถหาได้โดยวิธีเดียวกัน กล่าวคือ

$$n = \left(\frac{1.96 \times 0.507}{0.05}\right)^2 = 395 \text{ คน}$$

หรือคิดเป็นร้อยละ $\frac{395 \times 100}{1835} = 21.5$ ของนักศึกษาซึ่งเข้าศึกษาในปีการศึกษา 2523 ที่ไม่สำเร็จการศึกษา

เพื่อให้ผลการวิจัยมีความถูกต้องเชื่อถือได้มากขึ้น ผู้วิจัยได้สุ่มนักศึกษาที่สำเร็จการศึกษาแต่ละสาขาวิชามาร้อยละ 20 คิดเป็นจำนวนนักศึกษาทั้งหมด 530 คน สุ่มนักศึกษาที่ไม่สำเร็จการศึกษาแต่ละสาขาของทุกวิทยาลัยมาร้อยละ 25 คิดเป็นจำนวนนักศึกษาทั้งหมด 450 คน

สำหรับสาขานิเทศศาสตร์ ซึ่งเปิดสอนในมหาวิทยาลัยกรุงเทพเพียงแห่งเดียวและมีผู้ไม่สำเร็จการศึกษามากกว่าสาขาอื่น ๆ ได้คำนวณหาขนาดตัวอย่างตามวิธีข้างต้น ได้ตัวอย่างนักศึกษาที่ไม่สำเร็จการศึกษา 100 คน หรือประมาณร้อยละ 60 ของจำนวนนักศึกษาที่ไม่สำเร็จการศึกษาทั้งหมดและได้นักศึกษาที่สำเร็จการศึกษา 65 คน หรือประมาณร้อยละ 40 ของจำนวนนักศึกษาที่สำเร็จการศึกษาทั้งหมด

2. การสุ่มตัวอย่างใช้วิธีการสุ่มตัวอย่างแบบมีระบบโดยแบ่งชั้นภูมิ (Stratified Systematic Random Sampling) กล่าวคือ แบ่งนักศึกษาที่เข้าศึกษาในปีการศึกษา 2523 ออกเป็น 2 กลุ่มได้แก่ กลุ่มนักศึกษาที่สำเร็จการศึกษาและไม่สำเร็จการศึกษาแล้วในแต่ละกลุ่มของนักศึกษาที่สำเร็จและไม่สำเร็จการศึกษาแบ่งนักศึกษาออกตามวิทยาลัยและแต่ละวิทยาลัยแบ่งนักศึกษาออกตามสาขาวิชาที่ศึกษาจากนั้นจึงสุ่มตัวอย่างนักศึกษาจากแต่ละสาขาวิชาของวิทยาลัยต่าง ๆ ให้เป็นสัดส่วนกับจำนวนนักศึกษาทั้งหมดที่มีอยู่ในสาขาวิชาของคนนั้น ๆ ในอัตราร้อยละ 20 สำหรับผู้ที่สำเร็จการศึกษาและร้อยละ 25 สำหรับผู้ที่ไม่สำเร็จการศึกษา

3. ส่งแบบสอบถามที่เกี่ยวกับความคิดเห็นของนักศึกษาทางด้านต่าง ๆ ที่อาจมีผลกระทบต่อผลการศึกษาในระดับอุดมศึกษา ไปยังนักศึกษาตัวอย่างทางไปรษณีย์ เมื่อวันที่ 1 ธันวาคม 2527 โดยกำหนดให้นักศึกษาตัวอย่างส่งแบบสอบถามคืนภายในวันที่ 15 มกราคม 2528 สำหรับแบบสอบถามที่ไม่ได้รับคืนภายในกำหนด ผู้วิจัยได้เลือกตัวอย่างนักศึกษาเพิ่มเติมโดยให้นักศึกษา

ที่เลือกเพิ่มเติมนี้มีผลสัมฤทธิ์ทางการเรียนเมื่อสำเร็จการศึกษาหรือพันสภาพการศึกษาใกล้เคียงกับนักศึกษาที่ไม่ตอบแบบสอบถามภายในกำหนดมากที่สุด โดยเริ่มส่งแบบสอบถามรอบที่สองเมื่อวันที่ 22 มกราคม 2528 และกำหนดให้ส่งคืนภายในวันที่ 28 กุมภาพันธ์ 2528 สำหรับแบบสอบถามที่ไม่ได้รับกลับคืนมาในรอบสองนี้ ก็ทำเช่นเดียวกับรอบที่สองคือ เลือกตัวอย่างเพิ่มเติมแล้วส่งแบบสอบถามออกไปอีกในรอบที่สามระหว่างวันที่ 7 มีนาคม และกำหนดส่งคืนภายในวันที่ 15 เมษายน 2528 สำหรับแบบสอบถามที่ไม่ได้รับกลับคืนมาในรอบที่สามนี้ก็ทำเช่นเดียวกับรอบที่สองคือเลือกตัวอย่างเพิ่มเติมแล้วส่งแบบสอบถามออกไปอีกในรอบที่สี่ระหว่างวันที่ 20 เมษายนและกำหนดให้ส่งคืนภายในวันที่ 31 พฤษภาคม 2528 จำนวนตัวอย่างทั้งหมดที่เก็บรวบรวมได้จากการส่งแบบสอบถามทั้ง 4 รอบเท่ากับ 703 ราย คิดเป็นร้อยละ 71.73 ของจำนวนตัวอย่างทั้งหมด ในจำนวนนี้เป็นแบบสอบถามจากกลุ่มที่สำเร็จการศึกษา 350 ราย และกลุ่มที่ไม่สำเร็จการศึกษา 353 ราย รายละเอียดเกี่ยวกับจำนวนตัวอย่างนักศึกษาจำแนกตามสาขาวิชาและสถาบันการศึกษาแสดงไว้ในตารางที่ 2

ตารางที่ 2 จำนวนตัวอย่างนักศึกษาจำแนกตามสาขาวิชาการศึกษาและสถาบันการศึกษาได้ดังนี้

วิทยาลัยที่ศึกษา / สาขาวิชาที่ศึกษา	วิทยาลัยกรุงเทพ		วิทยาลัยหอการค้าไทย		วิทยาลัยธุรกิจบัณฑิต		วิทยาลัยเกริก		วิทยาลัยอัสสัมชัญ		วิทยาลัยเอเชียอาคเนย์		วิทยาลัยเทคนิคสยาม		รวมทุกวิทยาลัย	
	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ	สำเร็จ	ไม่สำเร็จ
บริหารธุรกิจ	26	34	31	31	24	29	16	23	43	17	33	23	40	19	213	176
บัญชี	17	23	14	18	18	14	13	14	14	14	7	9	3	2	86	94
นิเทศศาสตร์	51	83	-	-	-	-	-	-	-	-	-	-	-	-	51	83
รวมทุกสาขาวิชา	94	140	45	49	42	43	29	37	57	3.	40	32	43	21	350	353

2.3 การวิเคราะห์ข้อมูล

ในการวิเคราะห์ข้อมูล ผู้วิจัยได้นำข้อมูลที่เก็บรวบรวมได้มาทำการวิเคราะห์ตามระเบียบวิธีสถิติโดยใช้โปรแกรมสำเร็จรูป SPSS (Statistical Package for the Social Science) และแยกวิเคราะห์ดังนี้

2.3.1 การวิเคราะห์เกี่ยวกับตัวแปรด้านสถานภาพส่วนตัวของนักศึกษา

ศึกษาดูว่าการสำเร็จการศึกษาขึ้นอยู่กับเพศ สาขาที่เรียน ผลสัมฤทธิ์ทางการเรียนระดับมัธยมปลาย ผลสัมฤทธิ์ทางการเรียนระดับอุดมศึกษา ลำดับที่การสอบคัดเลือกเข้าในคณะวิชานั้น ๆ วุฒิสูงสุดเดิมก่อนเข้าศึกษาที่วิทยาลัย การศึกษาสูงสุดของบิดามารดา รายได้ของบิดา-มารดาหรือไม่ และศึกษาดูว่าตัวแปรใดใน 61 ตัวที่มีอิทธิพลกับเกรดเฉลี่ยสะสมระดับอุดมศึกษา (X_4) ซึ่งมีตัวแปร 61 ตัว ดังกล่าวคือ

X_1 = เพศ

X_2 = วิทยาลัยที่ศึกษา

X_3 = สาขาวิชา

X_4 = ระดับคะแนนเฉลี่ยสะสมในระดับอุดมศึกษา

X_5 = ลำดับที่การสอบคัดเลือกเข้าในคณะวิชานั้น ๆ

X_6 = วุฒิสูงสุดเดิมก่อนเข้าศึกษาในอุดมศึกษา

X_7 = ระดับคะแนนเฉลี่ยสะสมในระดับมัธยมปลาย

X_8 = สาเหตุที่ออกกลางคัน

X_9 = ผู้อุปการะการศึกษา

X_{10} = พักอาศัยอยู่กับใคร

X_{11} = สภาพความเป็นอยู่ของบิดามารดา

X_{12} = ผู้ที่เป็นหัวหน้าครอบครัวที่อาศัยอยู่

X_{13} = อาชีพบิดามารดาหรือผู้อุปการะ

X_{14} = รายได้ของบิดาและมารดาหรือผู้อุปการะ (บาท/เดือน)

X_{15} = การศึกษาสูงสุดของบิดามารดาหรือผู้อุปการะ

X_{16} = การเข้าชั้นเรียนสาย

X_{17} = ออกจากชั้นเรียนก่อนเวลาเลิก

X_{18} = ขาดเรียนบ่อย ๆ

X_{19} = ขาดเรียนบางวิชาเพื่อดูหนังสือสอบวิชาต่อไป

X_{20} = ไม่สนใจเนื้อหาที่เรียน

X_{21} = ไม่เตรียมตัวก่อนเข้าชั้นเรียน

X_{22} = ไม่ทบทวนหลังการเรียน

X_{23} = ไม่ส่งงานตามเวลาผู้สอนกำหนด

X_{24} = ไม่ศึกษาค้นคว้าจากหนังสืออื่น ๆ ที่เกี่ยวข้อง

X_{25} = ไม่เตรียมตัวในการสอบเท่าที่ควร

X_{26} = มีความเบื่อหน่ายต่อการเรียนทุกวิชา

X_{27} = ชอบหลีกเลี่ยงการทำงานที่ได้รับมอบหมาย

X_{28} = รู้สึกว่าจำเป็นต้องเรียนบางวิชาที่ไม่ชอบเลย

X_{29} = วิชาบังคับไม่น่าสนใจเท่าที่ควร

X_{30} = วิชาเลือกไม่น่าสนใจเท่าที่ควร

X_{31} = คิดว่าความรู้ที่คณะวิชาให้นั้นเป็นเรื่องแคบ ๆ

X_{32} = คิดว่าการศึกษาในคณะวิชานี้ให้ประโยชน์น้อย

X_{33} = รู้สึกว่าคณะวิชาที่ศึกษาด้อยกว่าคณะวิชาอื่น ๆ

X_{34} = รู้สึกว่าวิชาที่ศึกษายากเกินไปสำหรับความรู้ของท่าน

X_{35} = ข้อสอบยาก

X_{36} = วุฒิสูงสุดที่จบมา (ก่อนเข้ามาศึกษาในวิทยาลัย) ไม่เอื้ออำนวยสำหรับการศึกษาในคณะวิชานี้

X_{37} = ความรู้ความถนัดของท่านไม่ตรงกับคณะวิชาที่เรียน

X_{38} = ในขณะที่ศึกษาอยู่ท่านให้ความสนใจกับการทำกิจกรรมนักศึกษามากกว่าการเรียน

X_{39} = กิจกรรมเป็นสาเหตุสำคัญอันหนึ่งสำหรับท่านที่ทำให้ผลการเรียนไม่ดีเท่าที่ควร

X_{40} = ไม่สามารถแบ่งเวลาเรียนกับเวลาทำกิจกรรมแยกออกจากกัน

X_{41} = ท่านไม่ชอบทำกิจกรรม

X_{42} = ในขณะที่ศึกษาอยู่ท่านมักจดคำบรรยายไม่ทัน

X_{43} = ท่านไม่กล้าถามอาจารย์ผู้สอนเมื่อมีปัญหา

X_{44} = ไม่ค่อยเข้าใจที่อาจารย์บรรยาย

X_{45} = ไม่กล้าปรึกษากับอาจารย์ที่ปรึกษาเมื่อมีปัญหา

X_{46} = มีปัญหาเกี่ยวกับการทำแบบฝึกหัด

X_{47} = ค่าใช้จ่ายในการเรียนคณะวิชานี้

X_{48} = ต้องหาค่าใช้จ่ายในการเรียนด้วยตนเอง

X_{49} = ค่าใช้จ่ายสำหรับการศึกษาไม่เพียงพอ

X_{50} = ครอบครัวมีรายได้น้อยจึงไม่สามารถส่งเรียนจนจบหลักสูตรได้

X_{51} = ปัญหาในครอบครัวของท่านมีส่วนกระทบต่อผลการเรียน

X_{52} = บิดามารดาหรือผู้อุปการะไม่สนับสนุนการศึกษา

X_{53} = ต้องช่วยทางบ้านในการประกอบอาชีพ

X_{54} = มีภาระในการเลี้ยงครอบครัว

X_{55} = เกิดปัญหาส่วนตัวซึ่งมีผลกระทบต่อการเรียน

X_{56} = มีปัญหาสุขภาพทางกายหรือทางจิตที่เป็นอุปสรรคต่อการเรียน

X_{57} = ไม่สบายหนักจนทำให้ผลการเรียนต่ำลงมากเกินไป

X_{58} = การจัดครูเข้าชั้นเรียนไม่เหมาะสมกับความรู้และความถนัดในการสอน

X_{59} = ผู้สอนมีประสบการณ์ในการสอนน้อย

X_{60} = ผู้สอนไม่พยายามทำความเข้าใจปัญหาหรือความรู้สึกของนักศึกษา

X_{61} = ผู้สอนมีทัศนคติไม่ดีต่อผู้ที่เรียนอ่อน

X_{62} = ผู้สอนไม่มีเวลาที่จะให้นักศึกษาเข้าพบเพื่อซักถามปัญหา

การทดสอบดังกล่าวข้างต้นนี้จะใช้การทดสอบไคสแควร์ ซึ่งมีสมมติฐานของการทดสอบระหว่างเกรดเฉลี่ยสะสมระดับอุดมศึกษากับปัจจัยตัวแปรที่ i ดังนี้

H_0 : เกรดเฉลี่ยสะสมระดับอุดมศึกษาเป็นอิสระกันกับปัจจัยหรือตัวแปรที่ i

H_a : เกรดเฉลี่ยสะสมระดับอุดมศึกษาไม่เป็นอิสระกันกับปัจจัยสำหรับตัวแปรที่ i

ตัวสถิติที่ใช้ในการทดสอบคือ $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

เมื่อ χ^2 = ค่าสถิติไคสแควร์ ที่ยั้นความเป็นอิสระ

r = จำนวนแถว

c = จำนวนสดมภ์

O_{ij} = ค่าความถี่ที่สังเกตได้ ในแถวที่ i และสดมภ์ที่

E_{ij} = ค่าความถี่ที่คาดหวัง ในแถวที่ j และสดมภ์ที่

นำค่า χ^2 ที่คำนวณได้มาเปรียบเทียบกับค่า $\chi^2_{((r-1)(c-1),\alpha)}$ เมื่อ r และ c แทนจำนวนแถวและจำนวนสดมภ์ตามลำดับ ถ้า χ^2 คำนวณมากกว่า $\chi^2_{((r-1)(c-1),\alpha)}$ จะปฏิเสธ H_0 หรือยอมรับ H_a สำหรับการเปรียบเทียบว่าปัจจัยใดมีอิทธิพลต่อการสำเร็จการศึกษามากกว่ากัน ใช้ตัวสถิติคราแมร์วี ซึ่งมีสูตรในการคำนวณต่อไปนี้

$$V^2 = \frac{\chi^2}{n.Min(r-1,c-1)}$$

เมื่อ χ^2 คือ ค่าสถิติที่ได้จากการทดสอบความสัมพันธ์ระหว่างปัจจัยต่าง ๆ แต่ละปัจจัยกับลักษณะการสำเร็จการศึกษา

n คือ จำนวนนักศึกษาตัวอย่างที่นำมาทดสอบ

Min (r-1,c-1) คือ ค่าที่น้อยระหว่าง r-1 และ c-1 โดยที่ r และ c คือจำนวนแถวและจำนวนสดมภ์ของตารางการจรณ์ (Contingency table) ตามลำดับ

2.3.2 <u>การวิเคราะห์ตัวแปรเกี่ยวกับปัญหาส่วนตัว</u>

เปรียบเทียบความแตกต่างระหว่างกลุ่มที่สำเร็จการศึกษาและกลุ่มที่ไม่สำเร็จการศึกษาจำแนกตามหมวดปัญหา โดยใช้การทดสอบแบบที (t-test) ซึ่งมีสมมติฐานในการทดสอบสำหรับหมวดปัญหาที่ i ดังนี้

H_0 : คะแนนเฉลี่ยของหมวดปัญหาที่ i ของกลุ่มนักศึกษาที่สำเร็จและไม่สำเร็จไม่แตกต่างกัน

H_a : คะแนนเฉลี่ยของหมวดปัญหาที่ i ของกลุ่มนักศึกษาที่สำเร็จน้อยกว่ากลุ่มนักศึกษาที่ไม่สำเร็จ

ตัวสถิติที่ใช้ในการทดสอบคือ $t = \frac{(\bar{X}_{1i} - \bar{X}_{2i})}{\frac{S^2_{1i}}{n_{1i}} + \frac{S^2_{2i}}{n_{2i}}}$

ผู้วิจัยเลือกใช้สูตรนี้เนื่องจากไม่ทราบว่าค่าความแปรปรวนของคะแนนที่แท้จริงของทั้งสองกลุ่มเท่ากันหรือไม่

เมื่อ μ_{1i} และ μ_{2i} เป็นคะแนนเฉลี่ยจากประชากรของหมวดปัญหาที่ i ของนักศึกษากลุ่มที่สำเร็จและไม่สำเร็จตามลำดับ

$\bar{X}_{1i}$ และ $\bar{X}_{2i}$ เป็นคะแนนเฉลี่ยจากตัวอย่างของหมวดปัญหาที่ i ของนักศึกษากลุ่มที่สำเร็จและไม่สำเร็จตามลำดับ

S^2_{1i} และ S^2_{2i} เป็นค่าความแปรปรวนจากตัวอย่างของคะแนนของหมวดปัญหาที่ i ของนักศึกษากลุ่มที่สำเร็จและไม่สำเร็จตามลำดับ

n_{1i} และ n_{2i} เป็นขนาดของนักศึกษาตัวอย่างของหมวดปัญหาที่ i ของนักศึกษากลุ่มที่สำเร็จและไม่สำเร็จตามลำดับ

นำค่า t ที่คำนวณได้มาเปรียบเทียบกับค่า t ที่เปิดจากตารางด้วยระดับนัยสำคัญที่ต้องการ ซึ่งมีขั้นของความเป็นอิสระเท่ากับ

$$\frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\left[\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}\right]}$$

ถ้าค่า t ที่คำนวณได้มากกว่าค่า t ที่เปิดจากตาราง ที่ขั้นความเป็นอิสระข้างต้นแล้วระดับนัยสำคัญที่ต้องการจะปฏิเสธ H_0 หรือยอมรับ H_a

2.3.3 <u>การคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ของปัจจัย</u>

ระหว่างผลสัมฤทธิ์ทางการเรียนระดับมัธยมศึกษาตอนปลายและระดับปริญญาตรี จำแนกตามสาขาวิชา จำแนกตามกลุ่มการสำเร็จการศึกษา และรวมทุก ๆ สาขาวิชา คำนวณโดยใช้สูตร

$$r = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}}$$

เมื่อ r = ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

n = จำนวนนักศึกษาที่สำเร็จการศึกษาจากระดับอุดมศึกษาจากวิทยาลัยเอกชน

X_i = คะแนนเฉลี่ยสะสมระดับมัธยมศึกษา

Y_i = คะแนนเฉลี่ยสะสมระดับอุดมศึกษา

2.3.4 <u>ทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์</u>

ทดสอบสมมติฐานเกี่ยวกับผลสัมฤทธิ์ทางการเรียนของการศึกษาระดับมัธยมศึกษาตอนปลายและระดับปริญญาตรีว่ามีความสัมพันธ์กันหรือไม่ จำแนกตามสาขาวิชา และรวมทุกสาขา โดยทดสอบความมีนัยสำคัญของสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน ซึ่งมีสมมติฐานเพื่อการทดสอบเป็นดังนี้

H_0 : ผลสัมฤทธิ์ทางการเรียนระหว่างการศึกษาระดับมัธยมศึกษาและระดับปริญญาตรีไม่มีความสัมพันธ์กัน

H_a : ผลสัมฤทธิ์ทางการเรียนระหว่างการศึกษาระดับมัธยมศึกษาและระดับปริญญาตรีมีความสัมพันธ์กัน

หรือ $H_0 : \rho = 0$

$H_a : \rho \neq 0$

ตัวสถิติที่ใช้ในการทดสอบคือ $t_{n-1} = \frac{r}{S_r}$

เมื่อ $$r = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}}$$

$$S_r = \frac{1 - r^2}{\sqrt{n - 2}}$$

นำค่า t_{n-2} ที่คำนวณได้มาเปรียบเทียบกับค่า $t_{(n-2,\ \frac{\alpha}{2})}$ ถ้าค่า t ที่คำนวณมากกว่า $t_{(n-2,\ \frac{\alpha}{2})}$ จะปฏิเสธ H_0 หรือยอมรับ H_a

2.3.5 <u>การวิเคราะห์สหสัมพันธ์คาโนนิคอล</u> (Canonical Correlation Analysis)

ตามปกติเมื่อมีตัวแปรสุ่มอยู่ 2 ตัวคือ X และ Y จะสามารถหาความสัมพันธ์ของตัวแปรทั้งสองได้ โดยพิจารณาจากสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) หรือ ρ เมื่อ

$$\rho = \frac{Cov\ (X\ ,Y)}{\sqrt{Var(X)\ Var(Y)}}$$

ถ้าค่า Y ขึ้นอยู่กับค่า X หลาย ๆ ตัว วิธีการที่จะสามารถหาความสัมพันธ์ระหว่าง Y และ X เหล่านั้นจะใช้วิธีวิเคราะห์สหสัมพันธ์พหุคูณ (Multiple Correlation)

โดยให้ X เป็นปัจจัยหรือตัวแปรที่ต้องการศึกษาซึ่งมีอยู่ P ตัว

ดังนั้น $X' = [X_1\ X_2 \ldots\ldots X_p]$

และ $\hat{Y}$ เป็นค่าประมาณของผลรวมเชิงเส้น

$$\hat{Y}_j = \hat{\beta}_1 X_{1j} + \hat{\beta}_2 X_{2j} + \ldots\ldots + \hat{\beta}_p X_{pj}\ ,\ j = 1,2,\ldots,n$$

เมื่อ $\hat{\beta}_s$ (S = 1, 2,..., p) เป็นค่าสัมประสิทธิ์การถดถอยพหุคูณ (Multiple Regression Coefficient) ซึ่งหาค่าได้จากการประมาณโดยวิธี least-Square อันจะทำให้สามารถหาค่าสัมประสิทธิ์สหสัมพันธ์โดยการหาค่าความสัมพันธ์อย่างง่าย (Simple Correlation) ระหว่าง Y และ $\hat{Y}$ นั้นเอง ยังมีอีกกรณีหนึ่ง ซึ่งนอกจากจะมีตัวแปร X หลาย ๆ ตัวแล้วยังมี Y หลาย ๆ ตัวคือ

$$Y' = [Y_1\ Y_2\ \ldots Y_q]$$

วิธีการที่จะหาความสัมพันธ์ระหว่าง X หลาย ๆ ตัว และ Y หลาย ๆ ตัวก็คือ วิธีวิเคราะห์สหสัมพันธ์คาโนนิคอล จะได้ค่าความสัมพันธ์คาโนนิคอล (Canonical Correlation) ซึ่งหาได้จากความสัมพันธ์ระหว่างผลรวมเชิงเส้นของกลุ่มตัวแปร X และผลรวมเชิงเส้นของกลุ่มตัวแปร Y

เมื่อกำหนดค่า

$$X_j^* = \alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots\ldots + \alpha_p X_p \qquad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$
$$Y_j^* = \beta_1 Y_1 + \beta_2 Y_2 + \ldots\ldots\ldots + \beta_q Y_q$$

สมการนี้ยังไม่ทราบค่าของ α_i และ β_i ที่จะทำให้ความสัมพันธ์ระหว่าง Y*และ X* มีค่ามากที่สุดที่จะเป็นไปได้

ถ้าให้ ρ_c = Canonical Correlation Coefficient ระหว่าง X และY

ซึ่ง $$\rho_c = \frac{Cov(X^*, Y^*)}{\sqrt{Var(X^*)\ Var(Y^*)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

และ $\rho_c^{(1)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุด เรียกว่า First Canonical Correlation ซึ่งเกิดจาก X_1^* และ Y_1^*

$\rho_c^{(2)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสองเรียกว่า Second Canonical Correlation ซึ่งเกิดจาก X_2^* และ Y_2^*

$\rho_c^{(3)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสามเรียกว่า Third Canonical Correlation ซึ่งเกิดจาก X_3^* และ Y_3^*

. . . .
. . . .
. . . .
. . . .

X_1^* และ Y_1^* เรียกว่า first canonical variables

X_2^* และ Y_2^* เรียกว่า second canonical variables

X_3^* และ Y_3^* เรียกว่า third canonical variables

. .
. .
. .
. .

คู่ของ canonical variable ต่าง ๆ นั้น จะมีหลักการเหมือนกัน

กรณี $q < p$ ก็จะมี Canonical Correlation ทั้งหมด q ค่า

และ Canonical Variable ทั้งหมด q คู่

ผลลัพธ์ต่าง ๆ เหล่านี้ได้มาจากหลักการต่อไปนี้คือ

ถ้า X มีค่าเฉลี่ยเป็น μ_1 และความแปรปรวนเท่ากับ Σ_{11}

Y มีค่าเฉลี่ยเป็น μ_2 และความแปรปรวนเท่ากับ Σ_{22}

ให้ $$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]_{(p+q)\times(p+q)}$$

โดยที่ Σ_{11} เป็นเมตริกซ์ของความแปรปรวน (Variance matrix) ของ X ซึ่งมีขนาด p x p

Σ_{22} เป็นเมตริกซ์ของความแปรปรวน (Variance matrix) ของ Y ซึ่งมีขนาด q x q

Σ_{12} เป็นเมตริกซ์ของความแปรปรวนร่วม (Covariance matrix) ของ X และ Y ซึ่งมีขนาด p x q

$\Sigma_{21} = \Sigma_{12}'$

สามารถเขียนรายละเอียดได้เป็น

$$\Sigma = \left[\begin{array}{cccc:cccc} \sigma_{x_1x_1} & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_p} & \sigma_{x_1y_1} & \sigma_{x_1y_2} & \cdots & \sigma_{x_1y_q} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & \cdots & \sigma_{x_2x_p} & \sigma_{x_2y_1} & \sigma_{x_2y_2} & \cdots & \sigma_{x_2y_q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \sigma_{x_px_1} & \sigma_{x_px_2} & \cdots & \sigma_{x_py_p} & \sigma_{y_1y_1} & \sigma_{x_py_2} & \cdots & \sigma_{x_py_q} \\ \hdashline \sigma_{x_1y_1} & \sigma_{x_2y_1} & \cdots & \sigma_{x_py_p} & \sigma_{y_1y_1} & \sigma_{y_1y_2} & \cdots & \sigma_{y_1y_q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \sigma_{x_1y_q} & \sigma_{x_2y_q} & \cdots & \sigma_{x_py_q} & \sigma_{y_qy_1} & \sigma_{y_qy_2} & \cdots & \sigma_{y_qy_q} \end{array}\right]$$

Canonical variables ในรูปของ matrix คือ

$$\underset{\sim}{X}^* = X\,\underset{\sim}{\alpha}$$

$$\underset{\sim}{Y}^* = Y\,\underset{\sim}{\beta}$$

เมื่อ $\underset{\sim}{X}^*$ และ $\underset{\sim}{Y}^*$ เป็นเวคเตอร์ของ Canonical variable ขนาด p x 1 และ q x 1 ตามลำดับ

$\underset{\sim}{X}$ และ $\underset{\sim}{Y}$ เป็นเมตริกซ์ของตัวแปรสุ่มแรกเริ่ม (Original random variables) ขนาด p x p และ q x q ตามลำดับ

$\underset{\sim}{\alpha}$ เป็น Coefficient vectors ขนาด p x 1

$\underset{\sim}{\beta}$ เป็น Coefficient vectors ขนาด q x 1

ดังนั้นจากสมการ (2) สามารถเขียน Canonical Correlation Coefficient ได้เป็น

$$\rho_c = \frac{Cov(X\alpha, Y\beta)}{\sqrt{Var(X\alpha)\ Var(Y\beta)}}$$

หรือ

$$\rho_c = \frac{\alpha'\Sigma_{12}\beta}{\sqrt{(\alpha'\Sigma_{11}\alpha)(\beta'\Sigma_{22}\beta)}} \qquad \ldots\ldots\ldots \quad (2a)$$

หาค่า ρ_c จากสมการ (2a) โดยที่ต้องการได้ ρ_c ที่มีค่าสูงสุดดังนั้นจะ Maximize ρ_c เทียบกับ α และ β ซึ่งวิธีดังกล่าวนี้จะง่ายขึ้น เมื่อ Canonical Variable แต่ละตัวมีคุณสมบัติที่ว่าความแปรปรวนเป็น 1 นั่นคือ

$$Var\ (X_j^*) = \alpha'\Sigma_{11}\alpha = 1 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$Var\ (Y_j^*) = \beta'\Sigma_{22}\beta = 1 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

ซึ่งเงื่อนไขนี้จะทำให้ได้ $\rho_c = \alpha'\Sigma_{12}\beta$ เมื่อ Maximize ρ_c โดยวิธี Least Square ซึ่งจะได้ดังนี้

$$\Sigma_{12}\beta - \lambda\Sigma_{11}\alpha = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$\Sigma_{21}\alpha - \lambda\Sigma_{22}\beta = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

เมื่อ λ เป็น Lagrange multiplier

คูณสมการ (5) ด้วย λ ได้

$$\Sigma_{12}\lambda\beta = \lambda^2\Sigma_{11}\alpha \qquad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

คูณสมการ (6) ด้วย Σ_{22}^{-1}

$$\Sigma_{22}^{-1}\Sigma_{21}\alpha = \lambda\beta \qquad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

นำสมการ (8) แทนในสมการ (7)

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha = \lambda^2\Sigma_{11}\alpha$$

คูณทั้งสองข้างด้วย Σ_{11}^{-1} และจัดเทอมใหม่ จะได้

$$(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I)\alpha = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

ทำนองเดียวกันจะได้

$$(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I)\,\underset{\sim}{\beta} = 0 \quad \text{.....(10)}$$

ทั้งสมการ (9) และ (10) เป็น homogeneous equations ซึ่งสามารถหาคำตอบได้ แต่จากการที่ได้กำหนดไว้ก่อนแล้วว่า $q \leq p$ ดังนั้น ถ้าหากทราบค่า λ^2 และ β แล้วก็จะสามารถหา α ได้จากสมการ (5)

เนื่องจาก

$$\underset{\sim}{\alpha} = \frac{\Sigma_{11}^{-1}\Sigma_{12}\underset{\sim}{\beta}}{\lambda} \quad \text{.....(11)}$$

จากสมการ (5) เมื่อคูณด้วย α' จะได้

$$\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \lambda\underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha}$$

จากสมการ (6) เมื่อคูณด้วย β' จะได้

$$\underset{\sim}{\beta}'\Sigma_{21}\underset{\sim}{\alpha} = \lambda\underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta}$$

เนื่องจาก $\quad \underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha} = \underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta} = 1$

ดังนั้น $\quad \underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \underset{\sim}{\beta}'\Sigma_{21}\underset{\sim}{\alpha} = \lambda \quad \text{.....(12)}$

นั่นคือ จากสมการ (2a) จะได้ว่า $\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \lambda$ เป็น Canonical Correlation Coefficient ภายใต้เงื่อนไขของสมการ (3) และสมการ (4)

$$\rho_c = \lambda = \underset{\sim}{\alpha}'\Sigma_{12}\beta$$

ถ้าหากถอดรากที่สองของ characteristic root coefficient ที่มีค่ามากที่สุดของสมการ (9) หรือ (10) จะได้ first cannonical correlation

จากสมการ (10) หาค่า λ^2 และ β ได้ และจาก (11) หาค่า $\underset{\sim}{\alpha}$ ได้

ให้ $\lambda_1^2 > \lambda_2^2 \ldots\ldots > \lambda_q^2$ เป็น characteristic root ของสมการ (10)

$\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, \ldots\ldots, \underset{\sim}{\beta}_q$ เป็น characteristic vectors ที่สัมพันธ์กับ characteristic root λ_i^2 ดังกล่าว

$\alpha_1, \alpha_2, \ldots\ldots, \alpha_q$ เป็น characteristic vectors ซึ่งคำนวณได้จากการแทนค่า λ_i และ β_i ลงใน (11)

ดังนั้น Canonical Correlation Coefficient ที่ i คือ λi

และเซทของ ของ canonical variables ก็คือ $X_i^* = X\alpha_i$

$Y_i^* = Y\beta_i$

ค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอล (Canonical Correlation Coefficient) ที่ได้มีคุณสมบัติเช่นเดียวกับค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) กล่าวคือ ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลจะมีค่าอยู่ระหว่าง 0 และ 1 และจะไม่เปลี่ยนค่าแม้ว่าจะเปลี่ยนหน่วยการวัดก็ตาม

<u>การคำนวณ</u> เนื่องจากคุณสมบัติดังกล่าวข้างต้นของ Canonical Correlation Coefficient ดังนั้นเมื่อไม่ทราบค่า Σ เราจะประมาณ Σ ด้วย Covariance matrix(X) แทนโดยที่

$$S = \left[\begin{array}{c:c} S_{11} & S_{12} \\ \hdashline S_{21} & S_{22} \end{array}\right]$$

เมื่อ S เป็น Covariance Matrix ของตัวอย่างหรืออาจจะคำนวณได้จาก Correlation matrix (R) โดยที่

$$R = \left[\begin{array}{c:c} R_{11} & R_{12} \\ \hdashline R_{21} & R_{22} \end{array}\right]$$

เมื่อ R เป็น Correlation matrix ของตัวอย่าง

R_{11} เป็น matrix ขนาด p x p ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร X ซึ่งมี p ตัว

R_{22} เป็น matrix ขนาด q x q ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร Y ซึ่งมี q ตัว

R_{12} เป็น matrix ขนาด p x q ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร X และตัวแปร Y

$R_{21} = R_{12}'$

ค่าของ Canonical Correlation Coefficients ที่คำนวณจาก S หรือ R นั้น จะได้ผลเช่นเดียวกัน แต่ค่า α_i และ β_i นั้นจะได้ไม่เหมือนกัน กล่าวคือ

ถ้าหากคำนวณ α_i และ β_i จาก S matrix แล้ว ค่าที่ได้นั้นจะเป็นค่าตัวแปรแรกเริ่ม (Original Variables) ถ้าหากคำนวณจาก R matrix ค่าที่ได้จะเปลี่ยนอยู่ในรูปตัวแปรแรกเริ่มมาตรฐาน (Standardized Original Variables) อย่างไรก็ตามก็สามารถที่จะเปลี่ยนค่าสัมประสิทธิ์ที่คำนวณได้จากการใช้ R นั้นไปเป็นค่าสัมประสิทธิ์ที่ได้จาก S ได้ ดังจะเห็นได้จาก

$$\begin{bmatrix} \frac{1}{\sqrt{S_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{S_{22}}} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{S_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{S_{22}}} \end{bmatrix} = \begin{bmatrix} 1 & R_{12} \\ R_{21} & 1 \end{bmatrix}$$

ดังนั้น เมื่อนำ diagonal matrix ของส่วนกลับของค่าส่วนเบี่ยงเบนมาตรฐานมาคูณทั้งหน้าและหลังของ Covariance Matrix แล้ว ก็จะสามารถหาค่าของ Correlation matrix α, β ได้ กล่าวคือ

ถ้า $\alpha^*_{i1}, \alpha^*_{i2}, \ldots\ldots, \alpha^*_{ip}$ และ $\beta^*_{i1}, \beta^*_{i2}, \ldots\ldots \beta^*_{iq}$;

$(i = 1, 2\ldots\ldots, q)$

เป็นค่าที่คำนวณได้จาก matrix R แล้วจะสามารถหาค่า α, β ซึ่งคำนวณได้จาก S ได้โดยการหารแต่ละค่าสัมประสิทธิ์เหล่านั้นด้วยค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรแรกเริ่ม (Original Variable) นั่นคือ

$$\alpha^*_{i1}/_{S_{x_{11}}}, \quad \alpha^*_{i2}/_{S_{x_{22}}}, \quad \alpha^*_{i3}/_{S_{x_{33}}}, \quad \ldots\ldots\ldots, \alpha^*_{ip}/_{S_{x_{pp}}}$$

และ

$$\beta^*_{i1}/_{S_{y_{11}}}, \quad \beta^*_{i2}/_{S_{y_{22}}}, \ldots\ldots\ldots\ldots\ldots\ldots, \beta^*_{ip}/_{S_{y_{qq}}}$$

2.3.5.1 การทดสอบนัยสำคัญของการวิเคราะห์สหสัมพันธ์คาโนนิคอล

เมื่อทำการวิเคราะห์สหสัมพันธ์คาโนนิคอลแล้ว สิ่งที่จะได้คือ

(1) ρ_c

(2) α และ β

ค่าของ α และ β ซึ่งสอดคล้องกับ λ^2 ของเมตริกซ์ $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ โดยที่ λ^2 ก็คือความแปรปรวนที่ร่วมกันของตัวแปรทั้งสองชุดนั่นเอง ด้วยเหตุนี้ก็จะมี α และ β อยู่หลายชุดด้วยกัน ดังนั้นสิ่งที่จะต้องคำนึงถึงคือ Canonical Variables ทั้ง q คู่นั้นมีความสัมพันธ์กันหรือไม่ คือต้องมีการทดสอบสมมติฐานที่ว่า ตัวแปรทั้งสองชุดนั้นมีความสัมพันธ์กันหรือไม่นั่นคือทดสอบว่า Σ_{12} หรือ $\Sigma_{21} = 0$ หรือไม่

$H_0 : \Sigma_{12} = 0$ หรือ $H_0 : \Sigma_{21} = 0$

$H_a : \Sigma_{12} \neq 0$ $H_a : \Sigma_{21} \neq 0$

ใช้วิธีการทดสอบของ Bartlett ดังนี้

วิธีทดสอบของ Bartlett ใช้ตัวสถิติ V_0 โดยที่

$$V_0 = -\left[n-1-\frac{1}{2}(P+q+1)\right]\ln \Lambda_0$$

เมื่อ $$\Lambda_0 = \frac{1}{\prod_{i=1}^{q}\left[1+\frac{\lambda_i^2}{(1-\lambda_i^2)}\right]}$$

$$= \frac{1}{\prod_{i=1}^{q}\left[\frac{1}{(1-\lambda_i^2)}\right]}$$

$$= \prod_{i=1}^{q}(1-\lambda_i^2)$$

V_0 จะมีการกระจายแบบไคสแควร์ที่องศาความเป็นอิสระเท่ากับ pq ที่ระดับนัยสำคัญที่ต้องการ ถ้าค่า V_0 ที่คำนวณได้มากกว่าค่าไคสแควร์จากตาราง จะสรุปได้ว่า $\rho_c^{(1)}$ ซึ่งเป็น first canonical correlation coefficient นั้นไม่เท่ากับ 0 และจะทดสอบความสัมพันธ์ของ ρ_c ของคู่ที่เหลือต่อไปคือ

$$\Lambda_1 = \prod_{i=2}^{q} (1 - \lambda_i^2)$$

ซึ่งจะนำไปสู่ Barlett V_1 ดังนี้

$$V_1 = -\left[n-2-\frac{1}{2}(p+q+1)\right] \ln \Lambda_1$$

V_1 มีการกระจายแบบไคสแควร์ที่องศาความเป็นอิสระเท่ากับ (p-1)(q-1) การสรุปผลจะทำเหมือน V_0 ดังนั้น ถ้าหากมี k canonical correlation coefficients ที่ทดสอบแล้วปรากฏว่ามีนัยสำคัญส่วนที่เหลือก็จะถูกทดสอบว่า (q-k) Canonical Correlation Coefficient เหล่านั้นจะเท่ากับ 0 หรือไม่ กล่าวคือ

$$\Lambda_k = \prod_{i=(k+1)}^{q} (1 - \lambda_i^2)$$

ซึ่งจะได้ $V_k = -\left[n-1-k-\frac{1}{2}(p+q+1)\right] \ln \Lambda_k$

V_k มีการกระจายแบบไคสแควร์ที่มีองศาความเป็นอิสระเท่ากับ (p-k)(q-k)

ในการประมวลผลด้วยโปรแกรมสำเร็จรูป SPSS นั้น ผลลัพธ์ที่ได้จะมีนัยสำคัญหรือค่าของความน่าจะเป็นที่จะไม่ยอมรับ H_0 เมื่อ H_0 จริง ในกรณีที่จะทดสอบสมมติฐาน ณ ระดับนัยสำคัญ $\alpha = 0.05$ นั้น อาจใช้ค่าของความน่าจะเป็นที่จะไม่ยอมรับ H_0 เมื่อ H_o จริง มาเปรียบเทียบกับ $\alpha = 0.05$ เพื่อใช้เป็นเกณฑ์ในการตัดสินว่าจะยอมรับสมมติฐาน H_0 หรือไม่โดยที่

ถ้าค่าความน่าจะเป็นที่จะไม่ยอมรับ H_0 เมื่อ H_0 จริง < 0.05 จะไม่ยอมรับ H_0

ถ้าค่าความน่าจะเป็นที่จะไม่ยอมรับ H_0 เมื่อ H_0 จริง > 0.05 จะยอมรับ H_0

2.3.6 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

เป็นวิธีวิเคราะห์ทางสถิติโดยมีจุดมุ่งหมายที่จะคัดเลือกตัวแปรชุดหนึ่ง ซึ่งนักวิจัยคิดว่าตัวแปรชุดนี้มีความสัมพันธ์กับสิ่งที่ต้องการศึกษา จนถึงขั้นที่ตัวแปรชุดนี้เป็นตัวแบ่งแยกประชากรออกเป็นกลุ่มต่าง ๆ ได้อย่างชัดเจน

2.3.6.1 ขั้นตอนในการวิเคราะห์จำแนกประเภท มี 2 ขั้นตอนคือ

ขั้นที่ 1 การคัดเลือกตัวแปรชุดหนึ่ง เพื่อสร้างสมการที่ใช้ในการจำแนกประชากรออกเป็นกลุ่มต่าง ๆ กัน ได้อย่างชัดเจน สมการนี้คือ สมการจำแนกประเภท (Discriminant equation)

ขั้นที่ 2 การจำแนกตัวอย่างที่ได้ศึกษามานั้นเข้าเป็นสมาชิกของประชากรแต่ละกลุ่มโดยอาศัยสมการจำแนกประเภท

ในกรณีที่มีประชากร 2 กลุ่ม จะใช้วิธีการของ Fisher มาใช้ในการจำแนกประเภท กล่าวคือ ถ้ากำหนดให้

π_1 เป็นประชากรกลุ่มที่ 1

π_2 เป็นประชากรกลุ่มที่ 2

ตัวแปรที่ศึกษา คือ ตัวแปร X ซึ่งมีทั้งหมด p ตัวคือ

$$\underset{\sim}{X}' = [X_1\ X_2 \ldots\ldots\ldots\ldots X_p]$$

วิธีการของ Fisher นั้น จะแปลงตัวแปร X เหล่านี้ไปเป็นค่าของตัวแปรเพียงตัวเดียว คือ Y โดยที่ Y_1 และ Y_2 เป็นค่าสังเกตที่ได้จากประชากร π_1 และ π_2 ตามลำดับ

ถ้า μ_{1Y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร π_1

μ_{2Y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร π_2

$\left.\begin{array}{ll} \underset{\sim}{\mu}_1 = E(X/\pi_1) = \text{ค่าคาดหวังของตัวแปร X ที่มาจากประชากร } \pi_1 \\ \underset{\sim}{\mu}_2 = E(X/\pi_2) = \text{ค่าคาดหวังของตัวแปร X ที่มาจากประชากร } \pi_2 \end{array}\right\} \ldots(1)$

และโควาเรียนซ์เมตริกซ์ คือ

$$\Sigma = E(\underset{\sim}{X} - \underset{\sim}{\mu}_i)(\underset{\sim}{X} - \underset{\sim}{\mu}_i)' \quad ; \quad i = 1, 2, \ldots\ldots\ldots\ldots\ldots \quad (2)$$

พิจารณาผลรวมเชิงเส้น

$$\underset{(1x1)}{\underset{\sim}{Y}} = \underset{(1xp)}{\underset{\sim}{\beta}'} \underset{(px1)}{\underset{\sim}{X}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

เมื่อ $\underset{\sim}{Y}$ คือ Fisher's Linear Discriminant Function (ผลรวมเชิงเส้น)

$\underset{\sim}{\beta}$ คือ ค่าที่แสดงถึงความสำคัญของตัวแปร $X_1, X_2, \ldots, X_p$

ซึ่ง $\underset{\sim}{\beta}' = [\beta_1 \quad \beta_2 \ldots\ldots\ldots \beta_p]$

$\underset{\sim}{X}$ คือตัวแปรที่ต้องการศึกษา ซึ่งมีทั้งหมด p ตัว

และ $\underset{\sim}{X}' = [X_1 \quad X_2 \ldots\ldots\ldots X_p]$

ดังนั้น จะได้ $\mu_{1Y} = E(\underset{\sim}{Y}/\pi_1) = E(\underset{\sim}{\beta}'\underset{\sim}{X}/\pi_1) = \underset{\sim}{\beta}'\underset{\sim}{\mu}_1 \qquad \ldots\ldots\ldots (4)$

$\mu_{2Y} = E(\underset{\sim}{Y}/\pi_2) = E(\underset{\sim}{\beta}'\underset{\sim}{X}/\pi_2) = \underset{\sim}{\beta}'\underset{\sim}{\mu}_2$

และค่าความแปรปรวนของ Y จากทั้ง 2 ประชากรนั้นคือ

$$\sigma_Y^2 = \text{Var}(\underset{\sim}{\beta}'\underset{\sim}{X}) = \underset{\sim}{\beta}'\text{Cov}(\underset{\sim}{X})\underset{\sim}{\beta}$$

$$= \underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta} \qquad \ldots\ldots\ldots (5)$$

วิธีการของ Fisher คือพยายามหาผลรวมเชิงเส้นของค่า X ซึ่งจะทำให้ระยะทางระหว่าง μ_{1Y} และ μ_{2Y} มีค่ามากที่สุดที่จะเป็นไปได้

ค่าผลรวมเชิงเส้นซึ่งดีที่สุดจะต้องมีคุณสมบัติในการแบ่งแยกประชากรทั้ง 2 กลุ่มออกจากกันได้มากที่สุด ซึ่งวิธีการที่จะได้ผลรวมเชิงเส้นที่ดีนั้นก็โดยการหาค่า β ที่ทำให้อัตราส่วน $\frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ Y})^2}{\text{ความแปรปรวนของ Y}}$ มีค่ามากที่สุด

ให้ $\ell = \dfrac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ Y})^2}{\text{ความแปรปรวนของ Y}}$

$$= \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2}$$

$$= \frac{(\underset{\sim}{\beta}'\underset{\sim}{\mu}_1 - \underset{\sim}{\beta}'\underset{\sim}{\mu}_2)^2}{\underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}}$$

$$= \frac{\underset{\sim}{\beta}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{\beta}}{\underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}} \quad \ldots\ldots\ldots\ldots(6)$$

ค่า $\underset{\sim}{\beta}$ ที่ได้จะมีค่ามากที่สุดเมื่อ

$$\frac{\partial \ell}{\partial \underset{\sim}{\beta}} = 0$$

นั่นคือ $C(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \Sigma\underset{\sim}{\beta} = 0$

$$\underset{\sim}{\beta} = C\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

กำหนดให้ $C = 1$

จะได้ผลรวมเชิงเส้นเป็น

$$\underset{\sim}{Y} = \beta'\underset{\sim}{X} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X} \quad \ldots\ldots\ldots\ldots(7)$$

สมการ $\underset{\sim}{Y} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}$ สามารถใช้เป็นตัวจำแนกค่าสังเกตที่ได้มานั้นว่าจะอยู่ในประชากรกลุ่ม π_1 หรือ π_2 ได้โดยให้

$$Y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}_0$$

เป็นค่าของ discriminant function ของค่าสังเกต X_0 และให้

$$m = \frac{1}{2}(\underset{\sim}{\mu}_{1Y} + \underset{\sim}{\mu}_{2Y})$$

$$= \frac{1}{2}(\beta'\mu_1 + \beta'\mu_2)$$

$$= \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \quad \ldots\ldots\ldots\ldots(8)$$

เป็นจุดกึ่งกลาง (centriod) ระหว่างค่าเฉลี่ยของค่า Y จากทั้ง 2 ประชากร

สามารถเขียนกฎการจำแนกประเภทได้ดังนี้คือ

จัด X_0 ให้อยู่ใน π_1 ถ้าหากว่า $Y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}_0 < m$ .....(9)

จัด X_0 ให้อยู่ใน π_2 ถ้าหากว่า $Y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}_0 > m$

ในทางปฏิบัติจะไม่ทราบค่า $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ ดังนั้นถ้าสุ่มตัวอย่างจาก π_1 และ π_2 มาเป็นจำนวน n_1 และ n_2 แล้วโดยที่วัดค่าสังเกต

$$\underset{\sim}{X}' = \begin{bmatrix} X_1 & X_2 & \cdots\cdots\cdots\cdots & X_p \end{bmatrix}$$ จะได้ว่า

$$\underset{(pxn_1)}{\underset{\sim}{X}_1} = \begin{bmatrix} X_{11} & X_{12} & \cdots\cdots\cdots\cdots & X_{1n_1} \end{bmatrix}$$

$$\underset{(pxn_2)}{\underset{\sim}{X}_2} = \begin{bmatrix} X_{21} & X_{22} & \cdots\cdots\cdots\cdots & X_{2n_2} \end{bmatrix}$$

จะได้ค่าประมาณของ $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ เป็น $\bar{\underset{\sim}{X}}_1$, $\bar{\underset{\sim}{X}}_2$ และ $\underset{\sim}{S}_{pooled}^{-1}$ ตามลำดับโดยที่

$$\underset{(pxl)}{\bar{\underset{\sim}{X}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} \underset{\sim}{X}_{1j} \quad ; \quad \underset{(pxl)}{\bar{\underset{\sim}{X}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} \underset{\sim}{X}_{2j}$$

$$\underset{\sim}{S} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]\underset{\sim}{S}_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]\underset{\sim}{S}_2$$

$$= \frac{(n_1-1)\underset{\sim}{S}_1+(n_2-1)\underset{\sim}{S}_2}{n_1+n_2-2} \qquad \cdots\cdots\cdots\cdots\cdots(12)$$

เมื่อ $\underset{(pxp)}{\underset{\sim}{S}_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1} (\underset{\sim}{X}_{1j}-\bar{\underset{\sim}{X}}_1)(\underset{\sim}{X}_{1j}-\bar{\underset{\sim}{X}}_1)'$

$$\underset{(pxp)}{\underset{\sim}{S}_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2} (\underset{\sim}{X}_{2j}-\bar{\underset{\sim}{X}}_2)(\underset{\sim}{X}_{2j}-\bar{\underset{\sim}{X}}_2)'$$

ดังนั้น เมื่อแทนค่า $\bar{X}_1$, $\bar{X}_2$ และ $\underset{\sim}{S}_{pooled}^{-1}$ ลงใน (7) จะได้ Fisher's Sample Linear Discriminant Function เป็น

$$Y = \hat{\beta}'\underset{\sim}{X}$$

$$= (\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2)' \underset{\sim}{S}_{pooled}^{-1} \underset{\sim}{X} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots(13)$$

และค่าจุดกึ่งกลางระหว่างตัวแปร $\bar{Y}_1 = \hat{\beta}'\bar{X}_1$ และ $\bar{Y}_2 = \hat{\beta}'\bar{X}_2$ จะเขียนได้เป็น

$$\hat{m} = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$$

$$= \frac{1}{2}(\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} (\bar{X}_1 + \bar{X}_2) \quad \ldots\ldots\ldots\ldots\ldots(14)$$

กฎการจำแนกประเภทจะเขียนดังนี้

จัดให้ X_0 อยู่ใน π_1 ถ้า $Y_0 < \hat{m}$

จัดให้ X_0 อยู่ใน π_2 ถ้า $Y_0 \geq \hat{m}$

เมื่อ $Y_0 = (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} X_0$

จากสมการสามารถเขียนได้เป็น

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p$$

ค่า $\beta_1, \beta_2, \ldots\ldots \beta_p$ จะเป็นค่าแสดงถึงความสำคัญของตัวแปร $X_1, X_2, \ldots, X_p$ ซึ่งโดยปกติ X_i แต่ละค่านั้นจะมีหน่วยไม่เหมือนกัน การเปรียบเทียบค่า β_i เหล่านี้จะทำได้ก็โดยการปรับค่า β_i เหล่านี้ให้เป็นมาตรฐาน (Standardized) เสียก่อน

การแปลงค่า β_i เหล่านี้สามารถทำได้โดยเอาสมาชิกในแนวเส้นทแยงมุม (diagonal) ของเมตริกซ์ W มาถอดรากที่สอง แล้วนำไปคูณกับค่า β_i ทุก ๆ ค่าตามสูตรต่อไปนี้

$$\beta_i^* = (\sqrt{w_{ii}})(\beta_i)$$

เมื่อ W คือ $(X - \bar{X}_{g0})(X - \bar{X}_{g0})'$

$$X = \left[\begin{array}{cccc:cccc} x_{111} & x_{121} & \cdots & x_{1n1} & x_{112} & x_{122} & \cdots & x_{1n2} \\ x_{211} & x_{221} & \cdots & x_{2n1} & x_{212} & x_{222} & \cdots & x_{2n2} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ x_{p11} & x_{p21} & \cdots & x_{pn1} & x_{p12} & x_{p22} & \cdots & x_{pn2} \end{array}\right]$$

กลุ่มที่ 1 กลุ่มที่ 2

$$\underset{\sim}{\bar{X}}_{g0} = \begin{bmatrix} \bar{X}_{11} \cdots\cdots \bar{X}_{11} & \vdots & \bar{X}_{12} \cdots\cdots \bar{X}_{12} \\ \bar{X}_{21} \cdots\cdots \bar{X}_{21} & \vdots & \bar{X}_{22} \cdots\cdots \bar{X}_{22} \\ \vdots \qquad\qquad \vdots & \vdots & \vdots \qquad\qquad \vdots \\ \bar{X}_{p1} \cdots\cdots \bar{X}_{p1} & \vdots & \bar{X}_{p2} \cdots\cdots \bar{X}_{p2} \end{bmatrix}$$

β_i^* คือ Standardized discriminant weight ของ discriminant function

β_i คือ Row discriminant weight ของ discriminant function

W_{ii} คือ Diagonal element ของ W เมตริกซ์ โดยที่ i = 1, 2, ..., p

สมการจำแนกประเภทที่ได้มานั้นจะมีความสามารถในการแบ่งแยกกลุ่มได้ดีหรือไม่นั้น สามารถดูได้จากค่า miss classification error ซึ่งทราบได้จากการที่เปรียบเทียบว่าสิ่งที่เราศึกษาอยู่นั้นเป็นสมาชิกของประชากรกลุ่มหนึ่งและเมื่อใช้สมการจำแนกประเภทมาจัดการจำแนกแล้วกลับไปเป็นสมาชิกของอีกกลุ่มหนึ่งซึ่งทำให้ได้ผลที่ผิดจากความเป็นจริงไป

2.3.6.2 วิธีการคัดเลือกตัวแปรเข้าไปในสมการจำแนกประเภท (Selection of Variables)

จากการที่กล่าวมาแล้วข้างต้นนั้นเป็นหลักการโดยทั่ว ๆ ไปของวิธีการสร้างสมการจำแนกประเภทในกรณีที่หากว่าในการวิจัยนั้น มีตัวแปรเป็นจำนวนมาก การคัดเลือกตัวแปรให้เหลือจำนวนน้อยที่สุดแต่มีความสามารถในการใช้เป็นตัวจำแนกมากที่สุดนั้น เราสามารถทำได้โดยใช้การคัดเลือกตัวแปรทีละตัวโดยที่จะหาตัวแปรที่ดีที่สุดตัวแรกและตัวแปรที่ดีที่สุดตัวที่สอง ที่จะทำให้สมการจำแนกประเภทมีความสามารถจำแนกประเภทได้ดีที่สุด จากนั้นก็จะเลือกตัวที่สาม และตัวต่อไป ที่จะช่วยการจำแนกให้ดีขึ้นตามลำดับ ในแต่ละขั้นตอนตัวแปรที่ได้รับการคัดเลือกมาก่อนแล้วนั้น อาจถูกตัดทิ้งไปหากพบว่าเมื่อนำมารวมกับตัวแปรตัวอื่น ๆ แล้วไม่ช่วยให้สมการจำแนกประเภทได้ดีขึ้น วิธีการนี้เรียกว่าวิธีการสร้างสมการจำแนกประเภทแบบขั้นตอน (Stepwise discriminant analysis)

วิธีการคัดเลือกตัวแปรแบบขั้นตอนนี้เป็นวิธีการคัดเลือกตัวแปรอิสระทีละตัวตามอำนาจการจำแนกของตัวแปรนั้น โดยที่ตัวแปรตัวแรกที่ถูกเลือกเข้าไปนั้นจะมีอำนาจในการจำแนกสูงสุดเมื่อเทียบกับตัวแปรอื่น ๆ ตัวแปรที่เหลือก็จะถูกคัดเลือกเช่นเดียวกัน ทำเช่นนี้จนกระทั่งถึงเกณฑ์หนึ่ง ๆ ซึ่งเราได้กำหนดไว้ว่าตัวแปรที่เหลือนั้นจะไม่สามารถเข้าไปในสมการจำแนกได้อีกแล้ว วิธีการนี้จะคล้ายคลึงกับวิธีการวิเคราะห์ถดถอยพหุแบบขั้นตอน (Stepwise multiple regression analysis) ซึ่งโดยปกติเกณฑ์ที่ใช้คือ

กำหนดให้ $\hat{\ell}_p = (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2)$

เป็นค่าที่เกิดจากตัวแปร p ตัว

$\hat{\ell}_{p+q} = (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2)$

เป็นค่าที่เกิดจากตัวแปร p + q ตัว

จะกล่าวได้ว่าตัวแปร q ตัวที่เพิ่มขึ้นมานั้น จะไม่มีอำนาจในการจำแนกนั่นคือไม่ได้ช่วยให้สมการจำแนกเดิมที่ใช้ตัวแปรอยู่ p ตัวแล้วนั้นมีประสิทธิภาพดีขึ้น ถ้าหากว่า

$$F = \frac{n_1+n_2-p-q-1}{q} \times \frac{m(\hat{\ell}_{p+q} - \hat{\ell}_p)}{1 + m\hat{\ell}_p} < F_{\alpha,q,n_1+n_2-p-q-1}$$

โดยที่ n_1 = จำนวนตัวอย่างจากประชากร π_1

n_2 = จำนวนตัวอย่างจากประชากร π_2

$$m = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}$$

q = ค่าตัวแปรที่จะเข้าไปในสมการจำแนกในที่นี้โดยวิธีแบบขั้นตอน q จึงมีค่าเป็น 1 เสมอ

p = จำนวนตัวแปรที่มีอยู่เดิมแล้วในสมการจำแนก

α = ระดับนัยสำคัญที่กำหนดไว้

จากวิธีการดังกล่าวข้างต้น จะทำให้ได้สมการจำแนกประเภทและค่า eigenvalue ซึ่งเป็นค่าที่ได้จากขั้นตอนการหาสมการวิเคราะห์จำแนกประเภทโดยเป็นค่าความแปรปรวนของค่า Y ซึ่งได้จากการแปลงรูป (Transform) จากค่า X ต่าง ๆ ($X_1, X_2, \ldots\ldots, X_p$)

เป็นค่าที่ใช้วัดความสำคัญเชิงเปรียบเทียบของสมการที่ได้ โดยที่สมการวิเคราะห์จำแนกประเภทที่ได้มานั้นได้ตามลำดับความสำคัญของค่า eigenvalue ต่าง ๆ จากมากที่สุดและรองลงไปตามลำดับ ค่าผลรวมของ eigenvalue ทั้งหมดนั้นเป็นค่าของความแปรปรวน (Total variance) ทั้งหมดของตัวแปรจำแนกประเภท (ตัวแปร X) ค่า eigenvalue แต่ละค่าจึงคิดเป็นอัตราส่วนร้อยของค่ารวมของ eigenvalue ทั้งหมด ทำให้สามารถใช้ค่านี้ไปอ้างอิงถึงความสำคัญเชิงเปรียบเทียบของสมการจำแนกประเภทได้

2.3.6.3 <u>การทดสอบนัยสำคัญของสมการจำแนกประเภท</u>

เมื่อได้สมการจำแนกประเภทมาแล้วและอยากทราบว่าสมการนี้สามารถมีอำนาจในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติหรือไม่ เราสามารถทดสอบได้จาก

$$V_m = \left[N-1-(p+k)/2\right] \ln(1+\lambda_m)$$

โดยที่ V_m คือ Bartlett V Statistic ซึ่งมีการกระจายเป็นแบบ Chi-Square ที่มี d.f. = p+k-2m

N คือ จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

p คือ จำนวนตัวแปรทั้งหมด

k คือ จำนวนกลุ่ม

λ_m คือ Eigenvalues ต่าง ๆ เช่น ค่าตัวที่ 1, 2, 3.....,m

ค่า λ_m นี้คำนวณได้จาก

$$|A - \lambda_m I| = 0$$

เมื่อ A คือ โควาเรียนซ์ เมตริกซ์ของ X

ค่า λ_m นี้จะเลือกเอาเฉพาะค่าที่ไม่เป็นศูนย์

ถ้าค่า V_m ที่คำนวณได้มากกว่าค่า Chi-Square ที่ระดับนัยสำคัญที่กำหนดไว้แล้วแสดงว่าสมการนี้สามารถใช้จำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ

การคำนวณค่าอำนาจการจำแนกของกลุ่มของตัวแปร ((Total Discriminatory Power)

$$\hat{W}^2 = 1 - \frac{N}{(m-k)(1+\lambda_1)+(1+\lambda_2).....(1+\lambda_m)+1}$$

เมื่อ $\hat{w}^2$ = ค่าอำนาจในการแยกตัวแปรได้จากการประมาณค่า w^2

N = จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

k = จำนวนกลุ่ม

λ_m = Eigenvalues ต่าง ๆ

m = จำนวนค่า Eigenvalues

2.3.7 การวิเคราะห์ความถดถอยพหุคูณ (Multiple Regression Analysis)

ตัวแบบทางคณิตศาสตร์ที่จะนำมาใช้อีกวิธีหนึ่งที่จะอธิบายความสัมพันธ์ระหว่างสัมฤทธิ์ผลทางการเรียนกับสถานภาพส่วนตัวและหมวดปัญหาต่าง ๆ ที่เกี่ยวข้องกับตัวนักศึกษา ได้แก่ ตัวแบบเชิงเส้นของการถดถอยซึ่งเป็นที่นิยมใช้กันอย่างกว้างขวางในการวิเคราะห์เชิงเส้น ทั้งนี้เพราะส่วนใหญ่แล้วจะอยู่ในรูปของความสัมพันธ์ระหว่างตัวแปร โดยเฉพาะอย่างยิ่งถ้าหากว่าความสัมพันธ์อยู่ในรูปของความสัมพันธ์ระหว่างตัวแปรหลาย ๆ ตัว ดังเช่นที่เราได้ทดสอบ χ^2 แล้วว่า เกรดเฉลี่ยสะสมในระดับอุดมศึกษานั้นขึ้นอยู่กับ เพศ เกรดเฉลี่ยสะสมในระดับมัธยมศึกษาตอนปลายและปัจจัยอื่น ๆ อีก ในการวิเคราะห์เช่นนี้วิธีการถดถอยพหุคูณก็มีความเหมาะสมเพราะเราสามารถใช้ในการคำนวณเพื่อหาสมการทำนายใช้ในการสำรวจหาตัวแปรอิสระอีกครั้งหนึ่งหลังจากที่เราได้คัดตัวแปรโดยใช้ χ^2 แล้ว และนำตัวแปรที่ได้นี้มาคัดอีกทีโดยใช้วิธีการในการเลือกสมการถดถอยที่ดีที่สุด โดยการถดถอยพหุคูณนี้มีตัวแบบดังต่อไปนี้คือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + \varepsilon$$

โดยที่ β_i เป็นค่าสัมประสิทธิ์การถดถอยพหุคูณ ไม่ทราบค่าของตัวแบบ; $i = 0,1,2\ldots,k$

และถ้าหากว่าเรามีจำนวนข้อมูล n ตัว ต่อตัวแปรแต่ละตัวแล้วจะได้รูปของข้อมูลเป็นดังนี้คือ

ข้อมูล	ตัวแปรตาม	ตัวแปรอิสระที่ 1	ที่ 2	ที่ k
1	Y_1	X_{11}	X_{21}	X_{k1}
2	Y_2	X_{12}	X_{22}	X_{k2}
3	Y_3	X_{13}	X_{23}	X_{k3}
⋮	⋮	⋮	⋮	⋮
n	Y_n	X_{1n}	X_{2n}	X_{kn}

จากข้อมูลข้างต้นนี้ ถ้าหากเราจะนำมาเขียนในรูปของสมการเมตริกซ์จะได้เป็น

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots\cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots\cdots & X_{k2} \\ 1 & X_{13} & X_{23} & \cdots\cdots & X_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots\cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

หรือ $\underset{\sim}{Y} = \underset{\sim}{X}\beta + \underset{\sim}{\varepsilon}$

โดยที่ $\underset{\sim}{Y}$ คือข้อมูลของตัวแปรตาม โดยมีเวคเตอร์อันดับที่ nxl

$\underset{\sim}{X}$ คือข้อมูลของตัวแปรอิสระ โดยมีเมตริกซ์อันดับที่ nx(k+1)

$\underset{\sim}{\beta}$ คือสัมประสิทธิ์ของการถดถอย โดยมีเวคเตอร์อันดับที่ (k+1)xl

$\underset{\sim}{\varepsilon}$ คือค่าความคลาดเคลื่อนของค่าประมาณจากค่าจริงโดยมีเวคเตอร์อันดับที่ nxl

2.3.7.1 <u>ข้อสมมติเบื้องต้น</u>

ในการที่เราจะศึกษาถึงวิธีการประมาณและคุณสมบัติของรูปแบบซึ่งเราไม่ทราบค่านั้น เราจะต้องตั้งข้อสมมติ ซึ่งจำเป็นสำหรับการประมาณเสียก่อนและข้อสมมติที่จะกล่าวถึงต่อไปนี้ก็มีความสำคัญมาก เพราะคุณสมบัติของตัวประมาณโดยวิธีกำลังสองน้อยที่สุด (The Least Squares Method) ซึ่งเป็นวิธีที่เรานำมาใช้จะขึ้นอยู่กับข้อสมมติเหล่านั้น ซึ่งจะต้องนำมาพิจารณาประกอบก่อนลงมือทำการวิเคราะห์คือ

1. $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

นั่นคือ Y_i เป็นฟังก์ชั่นเชิงเส้นของ X_{ij} กับ ε_i โดยที่ $i = 1, 2, \ldots\ldots, n$

$j = 0, 1, \ldots\ldots, k$ และ $X_{oi} = 1$

2. $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$

นั่นคือ $E(\varepsilon_i) = 0$ สำหรับทุกค่า i

3. $E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I$

นั่นคือ $E(\varepsilon_i^2) = \sigma^2$ สำหรับทุกค่า i นั่นคือ ε_i มีความแปรปรวนคงที่

$E(\varepsilon_i \varepsilon_j) = 0$ เมื่อ $i \neq j$ แสดงว่าแต่ละคู่ของ ε ไม่เกี่ยวข้องกัน

(pairwise uncorrelated)

4. ค่าความคลาดเคลื่อนของค่าประมาณจากค่าจริง ε_i เป็นตัวแปรเชิงสุ่มที่มีการกระจายแบบปกติ $N(0,\sigma^2)$ ซึ่งข้อนี้จำเป็นต้องใช้ในกรณีที่ต้องการสร้างช่วงความเชื่อมั่นของ $\beta_0, \beta_1, \beta_2 \ldots\ldots \beta_k$ หรือเมื่อต้องการทดสอบสมมติฐานใด ๆ เกี่ยวกับตัวประมาณ

5. X เป็นเมตริกซ์อันดับที่ nx(k+1) ซึ่งมีค่าคงที่ จึงทำให้ X และ $\underset{\sim}{\varepsilon}$ เป็นตัวแปรอิสระกัน เราจึงได้ว่า

$$E(\underset{\sim}{\varepsilon}|X) = E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

$$E(X'\underset{\sim}{\varepsilon}) = X'E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

$$E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'|X) = E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I$$

6. X มี rank k+1 < n หมายความว่าจำนวนค่าสังเกต n จะมากกว่าจำนวนพารามิเตอร์ (k+1) ที่จะต้องประมาณค่าออกมา

จากข้อสมมติดังกล่าวมาแล้วข้างต้นเราจะเห็นได้ว่า

1. $E(\underset{\sim}{Y}|X) = X\underset{\sim}{\beta} + E(\underset{\sim}{\varepsilon}|X) = X\underset{\sim}{\beta}$

เช่น $E(Y_i|X_{1i}, X_{2i}, \ldots X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$

2. $E\left[(\underset{\sim}{Y}-E(\underset{\sim}{Y}))(\underset{\sim}{Y}-E(\underset{\sim}{Y}))'|X\right] = E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'|X) = \sigma^2 I$

เช่น $E\left[(Y_i-E(Y_i))(Y_i-E(Y_i))'|X_{1i}, X_{2i}, \ldots, X_{ki}\right] = \sigma^2$ สำหรับทุกค่า i

2.3.7.2 <u>การประมาณค่าพารามิเตอร์</u>

วิธีการในการประมาณค่าพารามิเตอร์การถดถอยนั้น กระทำได้ 2 วิธีคือ

1. วิธีกำลังสองน้อยที่สุด (Least Squares Method)
2. วิธีน่าจะเป็นสูงสุด (Maximum Likelihood Method)

ทั้ง 2 วิธีดังกล่าวเป็นวิธีในการคำนวณหาค่าเส้นถดถอยจากข้อมูลที่กำหนดให้หรือกล่าวอีกนัยหนึ่งก็คือ ถ้ากำหนดสมการ

$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ แล้วเทคนิคทั้งหมดข้างต้นก็จะเป็นเครื่องมือในการคำนวณ

$\hat{\underset{\sim}{Y}} = X\underset{\sim}{b}$ นั่นเอง หากแต่ว่าแต่ละวิธีก็มีคุณสมบัติของตัวมันเอง

แต่ในที่นี้เราจะนำมากล่าวถึงเพียงวิธีเดียวเท่านั้นคือวิธีที่ 1 ซึ่งถือว่าเป็นวิธีที่สำคัญและนิยมใช้กันมาก โดยมีหลักการของการประมาณแบบกำลังสองน้อยที่สุด คือการทำให้ผลรวมกำลังสองของส่วนเบี่ยงเบนของค่าสังเกตต่างจากค่าเฉลี่ยมีค่าน้อยที่สุดและจะทำให้ตัวประมาณที่ไม่เอนเอียง

จากรูปแบบ $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

และ $E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$ $\quad \because E(\underset{\sim}{\varepsilon}) = 0$

ซึ่ง $E(\underset{\sim}{Y})$ คือค่าเฉลี่ยของ $\underset{\sim}{Y}$

ดังนั้น $\underset{\sim}{\varepsilon} = \underset{\sim}{Y} - E(\underset{\sim}{Y})$

$= \underset{\sim}{Y} - X\underset{\sim}{\beta}$

โดยวิธีกำลังสองน้อยที่สุด เราจะต้องหาก่อนว่ากำลังสองของค่าความคลาดเคลื่อนเท่ากับ

$$\begin{aligned}\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} &= (\underset{\sim}{Y} - X\underset{\sim}{\beta})'(\underset{\sim}{Y} - X\underset{\sim}{\beta}) \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'X'\underset{\sim}{Y} - \underset{\sim}{Y}'X\underset{\sim}{\beta} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta} \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta}\end{aligned}$$

[เนื่องจาก $\underset{\sim}{Y}'X\underset{\sim}{\beta}$ เป็นสเกลล่า ดังนั้น ทรานสโพสของมันจึงมีค่าเท่าเดิม จึงได้ว่า

$(\underset{\sim}{Y}'X\underset{\sim}{\beta})' = \underset{\sim}{\beta}'X'\underset{\sim}{Y}$]

จากนั้นจึงเริ่มหาตัวประมาณโดยทำให้กำลังสองของค่าความคลาดเคลื่อนที่ได้มานั้นมีค่าน้อยที่สุดโดยใช้ การดิฟเฟอเรนชิเอท

$$\frac{\partial}{\partial\underset{\sim}{\beta}}(\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}) = \frac{\partial}{\partial\underset{\sim}{\beta}}(Y'Y - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \beta'X'X\underset{\sim}{\beta}) = 0$$

$$-2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{b} = 0$$

$$X'\underset{\sim}{Y} = X'X\underset{\sim}{b}$$

จากนี้จะหาตัวประมาณของ $\underset{\sim}{\beta}$ คือ $\underset{\sim}{b}$

ซึ่งเป็นชุดสมการปกติ (normal equations) และถ้าหากเราจะนำมาเขียนในรูปสเกลล่าจะได้ว่า

$$\Sigma Y_i = b_0 n + b_2\Sigma X_{1i} + b_2\Sigma X_{2i} + \cdots\cdots\cdots + b_k\Sigma X_{ki}$$

$$\Sigma X_{1i}Y_i = b_0\Sigma X_{1i} + b_1\Sigma X_{1i}^2 + b_2\Sigma X_{1i}X_{2i} + \cdots\cdots + b_k\Sigma X_{1i}X_{ki}$$

$$\Sigma X_{2i}Y_i = b_0\Sigma X_{2i} + b_1\Sigma X_{1i}X_{2i} + b_2\Sigma X_{2i}^2 + \cdots\cdots\cdots + b_k\Sigma X_{2i}X_{ki}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\Sigma X_{ki}Y_i = b_0\Sigma X_{ki} + b_1\Sigma X_{1i}X_{ki} + b_2\Sigma X_{2i}X_{ki} + \cdots\cdots + b_k\Sigma X_{ki}^2$$

การจะหา $\underset{\sim}{b}$ ได้ต้องมีเงื่อนไขว่าชุดสมการปกติทั้งหมด k+1 สมการนี้ต้องเป็นอิสระซึ่งกันและกันโดยที่ X'X จะต้องเป็นเมตริกซ์ที่ไม่เอกเทศ เพราะจะทำให้สามารถหา $(X'X)^{-1}$ ได้

$$\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$$

และเมื่อเราทำการดิฟเฟอเรนชิเอทอันดับที่ 2 ผลปรากฏว่า

$$\frac{\partial}{\partial\underset{\sim}{\beta}^2}(\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}) = 2X'X$$

ซึ่งเป็นเดฟฟินิทบวกเมื่อ X มี rank เต็มตามจำนวนคอลัมน์คือ k+1

เพราะฉะนั้น $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$ จึงเป็นตัวประมาณที่ได้จากการทำให้กำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด

จึงได้ว่า $\hat{\underset{\sim}{Y}} = X\underset{\sim}{b}$

ต่อไปเราจะพิจารณาค่าความแปรปรวนของ $\underset{\sim}{b}$

จาก
$$\begin{aligned}\underset{\sim}{b} &= (X'X)^{-1}X'\underset{\sim}{Y}\\ &= (X'X)^{-1}X'(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})\\ &= \underset{\sim}{\beta} + (X'X)^{-1}X'\varepsilon\end{aligned}$$

$$\underset{\sim}{b} - \underset{\sim}{\beta} = (X'X)^{-1}X'\underset{\sim}{\varepsilon}$$

$$\begin{aligned}E(\underset{\sim}{b} - \underset{\sim}{\beta})(\underset{\sim}{b} - \underset{\sim}{\beta})' &= E\left[(X'X)^{-1}X'\underset{\sim}{\varepsilon}\right]\left[(X'X)^{-1}X'\underset{\sim}{\varepsilon}\right]'\\ &= E\left[(X'X)^{-1}X'\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'X(X'X)^{-1}\right]\\ &= (X'X)^{-1}X'E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')X(X'X)^{-1}\\ &= \sigma^2(X'X)^{-1} \quad (\because E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I)\end{aligned}$$

เราจึงได้ค่า $\mathrm{Var}(\underset{\sim}{b}) = \sigma^2(X'X)^{-1}$

จากข้างต้นจึงได้ว่า $E(\underset{\sim}{b} - \underset{\sim}{\beta})(b - \beta')$ เป็น VAR - COV ของ $\underset{\sim}{b}$ แต่เนื่องจากไม่ทราบค่าของ σ^2 จึงต้องประมาณค่าของ σ^2 ก่อน

จาก $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

และ $\hat{\underset{\sim}{Y}} = X\underset{\sim}{b}$

ให้ $\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}}$

$= \underset{\sim}{Y} - X\underset{\sim}{b}$

$= \underset{\sim}{Y} - X(X'X)^{-1}X'\underset{\sim}{Y}$

$= (I - X(X'X)^{-1}X')\underset{\sim}{Y})$

$= M\underset{\sim}{Y}$ (ให้ $M = I - X(X'X)^{-1}X'$)

$= M(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$

$= MX\underset{\sim}{\beta} + M\underset{\sim}{\varepsilon}$

$= M\underset{\sim}{\varepsilon}$

ทั้งนี้เนื่องจาก $MX = 0$

โดย $MX = (I - X(X'X)^{-1}X')X$

$= X - X(X'X)^{-1}X'X$

$= 0$

หา $\underset{\sim}{e}'e = \underset{\sim}{\varepsilon}'MM\underset{\sim}{\varepsilon}$

$= \underset{\sim}{\varepsilon}'M\underset{\sim}{\varepsilon}$

ทั้งนี้เนื่องจาก $MM = (I-X(X'X)^{-1}X')(I-X(X'X)^{-1}X')$

$= I-X(X'X)^{-1}X'-X(X'X)^{-1}X'+X(X'X)^{-1}X'X(X'X)^{-1}X'$

$= I-X(X'X)^{-1}X'$

$= M$

$\therefore \underset{\sim}{e}'e = tr\underset{\sim}{\varepsilon}'M\underset{\sim}{\varepsilon}$

ทั้งนี้เนื่องจาก M เป็นสเกลล่าจึงเท่ากับ Trace ของตัวมันเอง

$\underset{\sim}{e}'e = tr\, M\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'$

$\therefore tr(AB) = tr(BA)$

$\therefore E(e'e) = E(trM\underset{\sim}{\varepsilon}\varepsilon')$

$= tr\ M.E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$

เพราะว่า trace เป็นฟังก์ชั่นเชิงเส้น

$$E(e'e) = \sigma^2 tr\ M \qquad ('.'\ E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I)$$

$$= \sigma^2 (n-k-1)$$

ทั้งนี้เนื่องจาก $tr\ M = tr(I - X(X'X)^{-1}X')$

$$= tr\ (I_n) - tr\ (X(X'X)^{-1}X')$$

$$= n - trX'X(X'X)^{-1} \qquad (tr(AB) = tr(BA))$$

$$= n - tr\ (I_{k+1})$$

$$= n-k-1$$

ดังนั้นจึงได้ว่า $S^2 = \dfrac{\underset{\sim}{e}'\underset{\sim}{e}}{n-k-1}$

$$= \dfrac{\underset{\sim}{Y}'M\underset{\sim}{Y}}{n-k-1}$$ ซึ่งเป็นตัวประมาณที่ไม่เอนเอียงของ σ^2

และ $S^2(X'X)^{-1}$ ก็เป็นตัวประมาณที่ไม่เอนเอียงของ $\sigma^2(X'X)^{-1}$

2.3.7.3 <u>คุณสมบัติของตัวประมาณโดยวิธีกำลังสองน้อยที่สุด</u>

จากข้อสมมติที่กำหนดให้ในหัวข้อข้างต้น ถ้าหากเป็นจริงแล้วเราจะได้ว่าการหาค่าตัวประมาณโดยวิธีกำลังสองน้อยที่สุดจะมีคุณสมบัติดังนี้คือ

1. $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$ เป็นตัวประมาณที่ไม่เอนเอียงของ $\underset{\sim}{\beta}$

พิสูจน์

$$\underset{\sim}{b} = (X'X)^{-1}X'Y$$

$$= (X'X)^{-1}X'(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$$

$$= (X'X)^{-1}X'X\underset{\sim}{\beta} + (X'X)^{-1}X'\underset{\sim}{\varepsilon}$$

$$= \underset{\sim}{\beta} + (X'X)^{-1}X'\underset{\sim}{\varepsilon}$$

$$E(\underset{\sim}{b}) = E(\underset{\sim}{\beta}) + (X'X)^{-1}X'E(\underset{\sim}{\varepsilon})$$

$$= \underset{\sim}{\beta}$$

นั่นคือ จะได้ว่า $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$ เป็นตัวประมาณที่ไม่เอนเอียงของ $\underset{\sim}{\beta}$

2. $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{Y}$ จะเป็นตัวประมาณที่ไม่เอนเอียงเชิงเส้นที่ดีที่สุด (Best Linear Unbiased Estimator) ในความหมายคือตัวประมาณอื่น ๆ ของ $\underset{\sim}{\beta}$ ซึ่งเป็นเชิงเส้นกับ $\underset{\sim}{Y}$ และไม่เอนเอียงเช่นกัน จะมีค่าวาเรียนซ์-โควาเรียนซ์เมตริก ซึ่งมีค่าเกินกว่าวาเรียนซ์-โควาเรียนซ์ของตัวประมาณ $\underset{\sim}{b}$ หรืออาจจะกล่าวได้อย่างง่าย ๆก็คือในบรรดาตัวประมาณในจำพวกของตัวประมาณที่ไม่เอนเอียงเชิงเส้นด้วยกันแล้วตัวประมาณโดยวิธีกำลังสองน้อยที่สุดนั้นจะมีค่าความแปรปรวนต่ำสุด โดยทฤษฎีของกอสล์-มาคอฟฟ์และเป็นที่นิยมใช้กันอย่างแพร่หลาย

จากการที่ตัวประมาณของ $\underset{\sim}{\beta}$ เป็นเชิงเส้นกับ Y เราจึงสามารถเขียนตัวประมาณนั้นได้เป็น $\underset{\sim}{A^*Y}$ โดย A^* เป็นเมตริกซ์ (k+1)xn ซึ่งไม่ขึ้นอยู่กับค่าของ $\underset{\sim}{Y}$

$$\text{ให้ } A^* = (X'X)^{-1}X' + A$$

$$\therefore A^*\underset{\sim}{Y} = \left[(X'X)^{-1}X' + A\right]\underset{\sim}{Y}$$

$$= \left[(X'X)^{-1}X' + A\right](X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$$

$$A^*\underset{\sim}{Y} = \left[I + AX\right]\underset{\sim}{\beta} + \left[(X'X)^{-1}X' + A\right]\underset{\sim}{\varepsilon}$$

$$E(A^*\underset{\sim}{Y}) = (I + AX)\underset{\sim}{\beta} + 0 \qquad \text{(จากข้อสมมติที่ 2)}$$

$$= \underset{\sim}{\beta} + AX\underset{\sim}{\beta}$$

ตัวประมาณ $A^*\underset{\sim}{Y}$ จะเป็นตัวประมาณที่ไม่เอนเอียง ถ้า $AX = 0$

$$\therefore \text{จาก } A^*\underset{\sim}{Y} = \underset{\sim}{\beta} + \left[(X'X)^{-1}X' + A\right]\underset{\sim}{\varepsilon}$$

$$A^*Y - \underset{\sim}{\beta} = \left[(X'X)^{-1}X' + A\right]\underset{\sim}{\varepsilon}$$

$$(A^*\underset{\sim}{Y} - \underset{\sim}{\beta})(A^*\underset{\sim}{Y}-\underset{\sim}{\beta})' = \left[(X'X)^{-1}X'+A\right]\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'\left[(X'X)^{-1}X'+A\right]'$$

$$E(A^*\underset{\sim}{Y}-\underset{\sim}{\beta})(A^*\underset{\sim}{Y}-\underset{\sim}{\beta})' = \left[(X'X)^{-1}X'+A\right]E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')\left[X(X'X)^{-1}+A'\right]$$

$$= \sigma^2\left[(X'X)^{-1}X'+A\right]\left[X(X'X)^{-1}+A'\right]$$

$$= \sigma^2\left[(X'X)^{-1}+AX(X'X)^{-1}+(X'X)^{-1}X'A'+AA'\right]$$

$$= \sigma^2\left[(X'X)^{-1}+AA'\right]$$

เนื่องจากเรากำหนดให้ $AX = 0 \rightarrow X'A' = 0$ ด้วย

จึงได้ว่า $Var(A^*Y) = \sigma^2(X'X)^{-1} + \sigma^2AA'$

$$= Var(\underset{\sim}{b}) + \sigma^2AA'$$

นั่นคือ วาเรียนซ์-โควาเรียนซ์ของตัวประมาณจะมีค่าเกิน วาเรียนซ์-โควาเรียนซ์ของตัวประมาณ ซึ่งประมาณโดยวิธีกำลังสองน้อยที่สุด

$\rightarrow \underset{\sim}{b} = (X'X)^{-1}X'Y$ จะเป็นตัวประมาณที่ไม่เอนเอียงเชิงเส้นที่ดีที่สุด

2.3.7.4 วิธีการทางสถิติที่ใช้ในการเลือกสมการถดถอยที่ดีที่สุด[1]

ในการเลือกสมการถดถอยเพื่อหาตัวแบบนั้น มีหลักเกณฑ์การเลือกที่เป็นหลักหรือข้อคิดที่สำคัญสำหรับการพิจารณาของนักวิจัยคือ ประการแรกต้องการให้ตัวแบบของสมการที่เลือกหรือที่สร้างขึ้นมานั้นมีประสิทธิภาพสูงและเป็นประโยชน์ในการพยากรณ์มากที่สุด ซึ่งในสมการนั้นจะประกอบด้วยตัวแปรตาม ซึ่งขึ้นอยู่กับตัวแปรอิสระมากที่สุด ดังตัวแบบเช่น

$$Y = f(X_1, X_2, \ldots\ldots X_k)$$

เพื่อว่าค่าของตัวแปรตามที่คำนวณได้จะเป็นค่าพยากรณ์ที่ใกล้เคียงและอีกประการหนึ่งเนื่องจากนักวิจัยทั้งหลายมักประสบกับปัญหาเกี่ยวกับเวลาและค่าใช้จ่ายในการวิเคราะห์ข้อมูล ยิ่งถ้าหากว่าการวิจัยนั้นใช้ตัวแปรอิสระมากเท่าไร (เพื่อที่จะให้มีประสิทธิภาพสูง) ก็ยิ่งจะทำให้ต้องสิ้นเปลืองค่าใช้จ่ายและเวลาในการคำนวณมากยิ่งขึ้นเท่านั้น

จากเหตุผลดังกล่าวมาแล้วข้างต้นจะเห็นได้ว่ามีความขัดแย้งในตัวมันเอง เพื่อที่จะเป็นการหาตัวแบบสมการถดถอยที่ดีที่สุด คือให้ได้ตัวแปรอิสระที่มีความเหมาะสมจำนวนน้อยแต่ให้สมการนั้นมีประสิทธิภาพสูงมีความเชื่อถือได้ จึงได้นำเอาวิธีการทางสถิติมาช่วยในการเลือกสมการถดถอยซึ่งกระทำได้หลายวิธี ในแต่ละวิธีนั้นก็มีเหตุผลในการเลือกตามความต้องการของแต่ละบุคคล และทั้งนี้ก็ขึ้นอยู่กับงบประมาณและความสะดวกด้วย ในที่นี้จะขอกล่าวถึงวิธีการทางสถิติในการเลือกสมการถดถอย ซึ่งมีอยู่หลายวิธีด้วยกันคือ

1. All Possible Regression
2. The Backward Elimination Procedure
3. The Forward Selection Procedure
4. The Stepwise Regression Procedure
5. Two Variations on the Four Previous Methods
6. The Stagewise Regression Procedure

---

[1] ดูรายละเอียดใน N.R. Draper and H. Smith, Applied Regression Analysis (New York: John Wiley & Sons, Inc., 1966), p 163-177.

วิธีการของการเลือกสมการถดถอยเพื่อจะดูถึงตัวแปรอิสระนี้เป็นเทคนิคทางสถิติวิธีหนึ่งที่ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม (dependent Variable Y) กับกลุ่มของตัวแปรอิสระ (independent variables $X_1, X_2, \ldots, X_k$) เพื่อที่จะเลือกว่าตัวแปรอิสระใดมีความสำคัญในขอบเขตที่มีนัยสำคัญ

ในการวิจัยครั้งนี้ เนื่องจากผู้วิจัยต้องการที่จะพิจารณาหาตัวแบบของสมการถดถอยที่เหมาะสม เพื่อที่จะดูว่าตัวแปรอิสระอะไรบ้างที่มีความสัมพันธ์พอที่จะอธิบายถึงสัมฤทธิ์ผลทางการเรียนของนักศึกษา จากการศึกษาเบื้องต้นได้ว่าตัวแบบอิสระ 62 ตัว ที่มีความสัมพันธ์กับสัมฤทธิ์ผลทางการเรียนโดยใช้ χ^2 ทำการทดสอบ ได้แก่ตัวแปรดังนี้

X_1 คือ เพศ

X_2 คือ วิทยาลัยที่ศึกษาอยู่

X_4 คือ G.P.A. ระดับอุดมศึกษา

X_7 คือ เกรดเฉลี่ยสะสมระดับมัธยมปลาย

X_{20} คือ ไม่สนใจเนื้อหาที่เรียน

X_{24} คือ ไม่ศึกษาค้นคว้าจากหนังสืออื่น ๆ ที่เกี่ยวข้อง

X_{33} คือ รู้สึกว่าคณะวิชาที่ศึกษาด้อยกว่าคณะวิชาอื่น ๆ

X_{34} คือ รู้สึกว่าวิชาที่ศึกษายากเกินไปสำหรับความรู้ของท่าน

X_{35} คือ ข้อสอบยาก

X_{41} คือ ท่านไม่ชอบทำกิจกรรม

X_{42} คือ ในขณะที่ศึกษาอยู่ท่านมักจดคำบรรยายไม่ทัน

X_{53} คือ ต้องช่วยทางบ้านในการประกอบอาชีพ

X_{55} คือ เกิดปัญหาส่วนตัวซึ่งมีผลกระทบต่อการเรียน

X_{58} คือ การจัดครูเข้าชั้นเรียนไม่เหมาะสมกับความรู้และความถนัดในการสอน

วิธีการถดถอยแบบเป็นขั้นตอน (The Stepwise Regression Procedure) เป็นวิธี ซึ่งผู้วิจัยได้ใช้ SPSS มาใช้ในการเลือกตัวแปรอีกครั้งหนึ่ง เพื่อจะได้ตัวแบบของสมการถดถอยครั้งนี้เนื่องจากเป็นวิธีที่รัดกุมและแก้ไขจุดบกพร่องของวิธีอื่น กล่าวคือในการตีงตัวแปรใหม่เข้ามาอยู่ในสมการถดถอยแต่ละครั้งนั้น จะมีการตรวจสอบดูว่าตัวแปรที่ดึงเข้ามาใหม่นี้จะมีอิทธิพลต่อตัวแปรที่อยู่ในสมการก่อนแล้วหรือไม่เพราะตัวแปรบางตัวที่เราเลือกเข้ามาอยู่ในตัวแบบในตอนแรกนั้นอาจไม่จำเป็น ถ้า

เราดูความสัมพันธ์ของมันกับตัวแปรที่เราดึงเข้ามาใหม่ ในการตรวจสอบดังกล่าวนี้จะใช้หลักของการทดสอบค่า F เพียงบางส่วน (Partial F-test) โดยถือว่าตัวแปรแต่ละตัวนั้นถูกใส่เข้าไปในตัวแบบเป็นตัวสุดท้าย ตัวแปรใดที่ไม่มีนัยสำคัญทางสถิติ (non-significance) ให้ตัดออกจากตัวแบบ คงไว้เฉพาะตัวที่มีนัยสำคัญ (Significance) วิธีการอันนี้จะหยุดต่อเมื่อเราไม่สามารถเพิ่มหรือลดตัวแปรใด ๆ ในตัวแบบได้อีก ซึ่งในที่นี้ผู้วิจัยจะขอกล่าวถึงขั้นตอนการเลือกตัวแปรโดยวิธีการถดถอยแบบเป็นขั้นตอนเพื่อให้ได้ตัวแบบที่ดีที่สุดโดยละเอียดดังนี้คือ

ขั้นที่ 1 คำนวณหาค่าของสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple correlation coefficient) ระหว่างตัวแปรอิสระทั้งหมด X กับตัวแปรตาม Y โดยใช้สัญญลักษณ์เป็น $r_{x,y}$ โดยที่

$$r_{x.y} = \frac{\sum_{i=1}^{n} (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i-\bar{X})^2 \sum_{i=1}^{n} (Y_i-\bar{Y})^2}}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - \frac{(\Sigma X_i)^2}{n})(\sum_{i=1}^{n} Y_i^2 - \frac{(\Sigma Y_i)^2}{n})}}$$

สัมประสิทธิ์สหสัมพันธ์นี้เป็นตัวที่ใช้ในการพิจารณาว่าตัวแปร 2 ชุดนั้นมีความสัมพันธ์กันหรือไม่อย่างไร โดยที่ถ้าตัวแปรชุดแรกแทนด้วย random variable X และตัวแปรชุดที่สองแทนด้วย random variable Y ค่าของความสัมพันธ์นี้จะอยู่ระหว่าง -1 ถึง 1 นั่นคือ

$$-1 \leq r_{xy} \leq 1$$

ถ้า $-1 \leq r_{x,y} < 0$ จะได้ว่า X กับ Y มีความสัมพันธ์กันในทางตรงข้าม นั่นคือ ถ้าหากค่าของ X มากขึ้นแล้วค่าของ Y จะลดลง หรือถ้าหากค่าของ X ลดลงแล้ว ค่าของ Y จะมากขึ้น

ถ้า $r_{X.Y} = 0$ จะได้ว่า X กับ Y นั้นไม่มีความสัมพันธ์กันเลย

ถ้า $0 < r_{X.Y} \leq 1$ จะได้ว่า X กับ Y มีความสัมพันธ์กันในทางเดียวกันนั่นคือถ้าหากค่าของ X มากขึ้นแล้ว ค่าของ Y จะมากขึ้นด้วย หรือถ้าหากค่าของ X ลดลงแล้ว ค่าของ Y จะลดลงด้วย

ซึ่งในการคำนวณนี้เนื่องจากเรามีตัวแปรอิสระหลายตัว ดังนั้นถ้าหากจะดูในรูปของเมตริกซ์แล้วจะสะดวกกว่าและได้เป็น

$$\begin{bmatrix} 1 & r_{X_1Y} & r_{X_2Y} \cdots\cdots\cdots r_{X_kY} \\ r_{YX_1} & 1 & r_{X_2X_1} \cdots\cdots\cdots r_{X_kX_1} \\ r_{YX_2} & r_{X_1X_2} & 1 \cdots\cdots\cdots r_{X_kX_2} \\ \vdots & \vdots & \vdots \\ r_{YX_k} & r_{X_1X_k} & r_{X_2X_k} \cdots\cdots\cdots 1 \end{bmatrix}$$

(เมื่อ k คือจำนวนตัวแปรอิสระที่เรานำมาพิจารณา)

จากนี้จะดูว่าตัวแปรอิสระใดที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตามสูงสุดก็จะได้รับเลือกเข้าสมการถดถอยก่อนในที่นี้สมมติว่า r_{X_1Y} มีค่าสูงสุด ดังนั้น X_1 จึงเป็นตัวแปรอิสระที่จะเลือกมาพิจารณาสร้างสมการก่อนได้สมการถดถอยเป็น

$$\hat{Y} = b_0 + b_1X_1$$

จากสมการนี้จะนำมาทดสอบดูว่าตัวแปรอิสระ X_1 ซึ่งนำเข้ามาอยู่ในสมการนั้นจะมีส่วนช่วยในการอธิบายค่าของตัวแปรตาม อย่างมีนัยสำคัญทางสถิติหรือไม่ โดยการใช้ F-test ดูค่า F เทียบกับ F จากตารางในระดับนัยสำคัญที่กำหนดไว้ |ในที่นี้เทียบกับ $F(1, n-1, 1-\alpha)$ ในตารางโดย n คือ จำนวนข้อมูล, α คือ ระดับนัยสำคัญที่กำหนด| ถ้าค่า $F \geq F$ จากตาราง แสดงว่า X_1 ยังคงอยู่ในสมการ แต่ถ้าหากว่า

Sequential F $\leq$ F จากตารางแสดงว่า Y ไม่มีความสัมพันธ์กับตัวแปรอิสระใดอย่างมีนัยสำคัญเลย ก็จะหยุดอยู่เพียงแค่นี้

ในการทดสอบนี้จะเห็นได้ว่า ในการทดสอบเรากล่าวว่าทดสอบ ซึ่งอันที่จริงแล้วคือการทดสอบ F ของ Regression Coefficient ของตัวแปรอิสระที่กำลังพิจารณาอยู่นั่นเอง อย่างในกรณีนี้สมมติว่า X_1 คือตัวแปรอิสระที่ถูกเลือกเข้าในสมการ ดังนั้นจึงทดสอบสมมติฐาน $H_0 : \beta_1 = 0$

<u>ขั้นที่ 2</u> คำนวณหาค่าของสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วน (Partial Correlation Coefficient) ระหว่างตัวแปรอิสระที่เหลือ X_j $(j \neq 1)$ กับตัวแปรตาม Y โดยกำหนดให้ตัวแปรอิสระ X_1 ซึ่งเราเลือกไว้ในสมการถดถอยแล้วคงที่และใช้สัญญลักษณ์เป็น $r_{X_jY.X_1}$ โดยที่

$$r_{X_jY.X_1} = \frac{r_{X_jY} - r_{X_jX_1} \cdot r_{YX_1}}{\sqrt{(1 - r^2_{X_jX_1})(1 - r^2_{YX_1})}}$$

ตัวแปรอิสระที่เหลือ X_j ตัวใดที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนกับตัวแปรตาม Y สูงสุด จะได้รับเลือกเข้าสมการต่อไป สมมติว่าในที่นี้ $r_{X_2Y.X_1}$ ให้ค่าสูงสุด ดังนั้น X_2 จะเป็นตัวแปรอิสระที่เลือกมาพิจารณาและสมการถดถอยจะเป็น

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

จากสมการนี้จะมีการทดสอบ F ของ X_2 ดูค่า X_2 ไม่สำคัญพอก็จะตัด X_2 ออกเหลือ $\hat{Y} = b_0 + b_1X_1$ เช่นเดิมและหยุดอยู่เพียงแค่นี้ แต่ถ้า X_2 สำคัญแล้วแสดงว่าต้องคงสมการ $\hat{Y} = b_0 + b_1X_1 + b_2X_2$

<u>ขั้นที่ 3</u> เราจะทดสอบ partial F ของตัวแปรอิสระที่เลือกเข้าสมการก่อนหน้านี้ ในที่นี้คือ X_1 นั่นคือ ดูว่าหลังจากที่เลือก X_2 เข้ามาเพิ่มในสมการแล้ว X_1 ซึ่งเป็นตัวแปรอิสระ

ซึ่งเลือกเข้ามาในสมการ ก่อนจะยังมีความสำคัญอยู่อีกหรือไม่ โดยจะทดสอบ Partial F ของ X_1 เมื่อ X_2 อยู่ในสมการเรียบร้อยแล้วเทียบกับ F จากตาราง $[F(1/n-3,1-\alpha)]$ ถ้าหากว่าค่า partial F $\leq$ F จากตารางแล้วจะได้ว่า X_1 ซึ่งอยู่ในสมการตอนแรกนั้นจะไม่มีความสำคัญอีกต่อไป ดังนั้นจะตัด X_1 ออกสมการที่ได้จะเปลี่ยนใหม่เป็น $\hat{Y} = b_0 + b_2X_2$ แต่ถ้า partial F > F จากตารางแล้วแสดงว่า สมการ $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ จะยังคงใช้ได้ และเราจะเริ่มหาตัวแปรอิสระตัวใหม่ต่อไป

จากขั้นตอนนี้จะพบว่ามีการทดสอบค่า Partial F ซึ่งในที่นี้ก็คือการทดสอบดูว่าตัวแปรอิสระที่เข้าไปอยู่ในสมการแล้ว ถ้าให้เข้าเป็นตัวสุดท้ายโดยให้ตัวแปรอิสระตัวอื่นที่เลือกแล้วอยู่ในสมการก่อนจะสามารถอธิบายค่า Y ได้อีกมากน้อยเพียงใด

<u>ขั้นที่ 4</u> จากตอนนี้เราได้แล้วว่า X_1, X_2 เป็นตัวแปรอิสระที่เลือกอยู่ในสมการแล้ววิธีการต่อไปคือ จะย้อนกลับไปทำเช่นเดียวกับขั้นที่ 2 อีกคือหาค่าของสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนระหว่างตัวแปรอิสระที่เหลือ X_j $(j \neq 1, 2)$ กับตัวแปรตาม Y โดยกำหนดให้ตัวแปรอิสระ X_1, X_2 ซึ่งเราเลือกไว้แล้วในสมการถดถอยนั้นคงที่และใช้สัญญลักษณ์เป็น $r_{X_jY.X_1X_2}$ โดยที่

$$r_{X_jY.X_1X_2} = \frac{r_{X_jYX_2} - r_{X_jX_1.X_2}\, r_{YX_1.X_2}}{\sqrt{(1 - r^2_{X_jX_1.X_2})(1 - r^2_{YX_1.X_2})}}$$

หรือ

$$= \frac{r_{X_jYX_1} - r_{X_jX_2.X_1}\, r_{YX_2.X_1}}{\sqrt{(1 - r^2_{X_jX_2.X_1})(1 - r^2_{YX_2.X_1})}}$$

ตัวแปรอิสระใดที่ให้ค่าของสัมประสิทธิ์สหสัมพันธ์เพียงบางส่วนสูงสุดก็จะได้รับการเลือกเป็นตัวแปรอิสระเข้าสมการตัวต่อไปและจะทำการทดสอบเช่นเดียวกับที่ผ่านมาจนกว่าจะไม่มีตัวแปรอิสระใดที่เป็นที่ยอมรับคือไม่ผ่านการทดสอบ F test ก็จะได้สมการอันสุดท้ายเป็นอันที่ต้องการ