การทำ Multiple Classification Analysis จาก Dummy Variable Regression*

มีงสรรพ์ สันติกาญจน์

ภาควิชาเศรษฐศาสตร์ คณะสังคมศาสตร์

มหาวิทยาลัยเชียงใหม่

บทนำ

Multiple Classification Analysis หรือเรียกย่อ ๆ ว่า MCA เป็นวิธีวิจัยทางสถิติสำหรับสังคมศาสตร์ ซึ่งเริ่มใช้เมื่อปี ๑๙๖๓ ต่อมาก็มีการใช้อย่างแพร่หลายในสังคมศาสตร์หลายสาขา เช่น ประชากรศาสตร์ สังคมวิทยา และการวิจัยทางเศรษฐศาสตร์ ซึ่งมีตัวแปรทางสังคมเข้ามาเกี่ยวข้อง บทความนี้ มิได้มีจุดประสงค์ที่จะเผยแพร่วิธีทางสถิตินี้ เพราะมีผู้รู้จักและสามารถอาศัย Program ของ MCA โดยตรงหรือ ANOVA Program ของ SPSS ได้ แต่ต้องการแสดงถึงความสัมพันธ์ระหว่าง MCA กับ Multiple Regression Analysis (MRA) ซึ่งมีตัวแปรอิสระเป็นตัวแปร dummy และการหาสัมประสิทธิ์สำหรับ MCA จาก MRA ทั้งนี้จะมีประโยชน์มากสำหรับนักวิจัย ในมหาวิทยาลัยต่างจังหวัดซึ่งไม่มี MCA Program หรือ SPSS ให้ใช้ หรือสำหรับการวิจัยซึ่งมีตัวแปรมากเกินกว่าที่ ANOVA program ใน SPSS จะรับไว้ได้ ในตอนต้นของบทความนี้จะกล่าวถึงความหมายของ MCA ตอนที่ 2 จะเปรียบเทียบโมเดล MCA กับโมเดลของ MRA ในตอนสุดท้ายจะให้ตัวอย่างการเปลี่ยนสัมประสิทธิ์ MRA เป็นสัมประสิทธิ์ MCA

ประโยชน์ของ MCA

MCA ใช้ตอบปัญหาได้เช่นเดียวกับ MRA เช่นสามารถบอกได้ว่า ตัวแปรอิสระทั้งหมดและตัวแปรแต่ละตัวหรือกลุ่ม สามารถอธิบายการเปลี่ยนแปลงของตัวแปรตามได้ดีเพียงใด และ

* ผู้เขียนขอขอบคุณมูลนิธิ Ford และ Rockefeller ซึ่งได้ให้โอกาสผู้เขียนเข้าร่วมการประชุมเชิงปฏิบัติการเรื่อง Multivariate Data Analysis ที่ AIT ระหว่างวันที่ ๒๘ มิถุนายน ถึงวันที่ ๑๘ สิงหาคม ๒๕๒๒ ผู้เขียนได้รับประโยชน์จากการแนะนำวิธีวิจัยทางสถิตินี้จาก ศาสตราจารย์ R.W. Hodge สูตร สมการต่าง ๆ ตลอดจนการพิสูจน์ความสัมพันธ์ในภาคผนวกนั้นได้มาจากการบรรยายของศาสตราจารย์ Hodge ในการประชุมดังกล่าว แต่ข้อผิดพลาดใด ๆ ในบทความนี้ผู้เขียนขอยอมรับว่าเป็นความผิดของตนเองทั้งสิ้น

สามารถใช้คาดคะเน (predict) ขนาดของตัวแปรตามเมื่อขนาดของตัวแปรอิสระเปลี่ยนไป นอกจากนั้นยังสามารถยืนยันว่า ความสามารถในการคาดคะเนนั้น มีความเป็นไปได้ทางสถิติเท่าใด

MCA สามารถใช้ได้ เมื่อความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามไม่เป็นความสัมพันธ์เชิงเส้นตรง (non-linear relationship) เมื่อตัวอิสระมีความสัมพันธ์กันเอง (correlated) และเมื่อตัวแปรอิสระเป็นตัวแปรประเภทกลุ่ม (nominal scale) ตัวอย่างเช่น ในการหาผลกระทบของรายได้ อาชีพของสามี และศาสนาต่อจำนวนบุตรเกิดรอดนั้น ตัวแปรตามคือจำนวนบุตร และรายได้ซึ่งเป็นตัวแปรอิสระเป็นตัวแปรที่เป็น interval scale ส่วนตัวแปรอิสระที่เป็นตัวแปรประเภทกลุ่มได้แก่ตัวแปรอาชีพ และศาสนาซึ่งอาจจะแบ่งย่อยออกเป็นชุดละ 4 กลุ่ม (classification) เช่น กลุ่มอาชีพต่าง ๆ ได้แก่ข้าราชการ พ่อค้า ชาวไร่ชาวนาและอื่น ๆ กลุ่มศาสนาได้แก่ศาสนาพุทธ คริสต์ อิสลาม และการนับถือผี เป็นต้น

ตัวอย่างเช่น ในการหาผลกระทบของชีวปัจจัยและปัจจัยเศรษฐกิจสังคมต่อภาวะเจริญพันธุ์ สามารถชี้ให้เห็นถึงอิทธิพลของตัวแปรต่าง ๆ ได้จากสัมประสิทธิ์ที่แสดงถึงความบ่ายเบนจากค่าตัวกลางอันเนื่องจากอิทธิพลของตัวแปรโดยได้คำนึงถึงอิทธิพลของตัวแปรอื่น จากสดมภ์ที่ 4 ตารางที่ 1 ซึ่งอธิบายได้ว่า ตามปกติแล้ว สตรีที่สมรสแล้วในชนบทอำเภองาวจะตั้งครรภ์โดยเฉลี่ย 3.61 ครั้ง แต่ถ้าสตรีผู้นั้นอายุ 20 ปี หรือต่ำกว่าจะตั้งครรภ์เพียง $3.61 - 2.99 = .62$ ครั้ง หากสตรีคนเดียวกันนั้นมีสามีที่ทำการกสิกรรมก็จะมีบุตรเพียง $.62 - 0.07 = .55$ ครั้ง สตรีอายุ 41 ปีขึ้นไป ซึ่งมีสามีรับราชการและมีรายได้ของครอบครัวต่อหัวต่อปี 4,501 บาทขึ้นไปจะตั้งครรภ์ $3.61 + 1.65 - 0.07 - 0.66$ เท่ากับ 4.53 ครั้ง ดังนั้น เราสามารถเลือกเอาอิทธิพลของตัวแปรที่เราสนใจเป็นพิเศษมาแสดงเป็นแผนภาพดังแผนภาพที่ 1

ค่าของ eta นั้น แสดงถึงความสามารถของตัวแปรอิสระแต่ละตัวหรือชุด (classifications) ซึ่งได้รวมทุกกลุ่มที่จะอธิบายความเปลี่ยนแปลงตัวแปรตาม eta^2 ก็คือสัดส่วนของ Total sum of square ที่อธิบายได้โดยตัวแปรอิสระ ส่วน beta นั้น วัดความสามารถของตัวแปรอิสระตัวหนึ่งในการอธิบายการแปรเปลี่ยนของตัวแปรตามโดยที่ได้คำนึงถึงอิทธิพลของตัวแปรอื่น ๆ

การเปรียบเทียบสัมประสิทธิ์ MCA กับ MRA

อันที่จริงแล้ว MCA ก็คือ MRA ที่มีตัวแปรตามที่เป็นตัวแปรประเภท dummy นั่นเอง มีข้อแตกต่างเพียงเล็กน้อยตรงที่สัมประสิทธิ์ของ MCA แสดงถึงความบ่ายเบนจากค่าตัวกลาง (Grand

ตารางที่ 1 ปัจจัยที่กำหนดภาวะเจริญพันธุ์ของสตรีสมรสแล้วในชนบทอำเภองาว จังหวัดลำปาง เมื่อ พ.ศ. 2522

Grand mean = 3.61

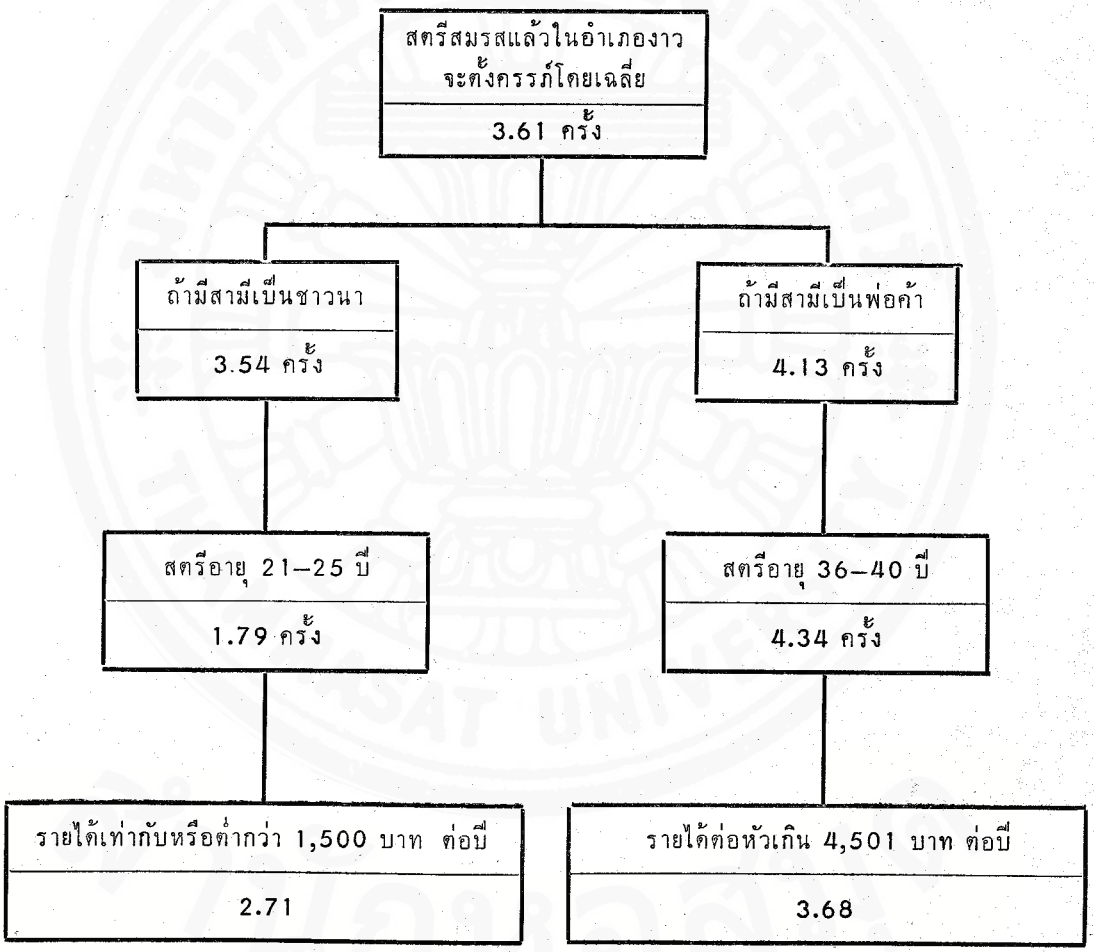
ปัจจัยที่พิจารณา	(I) จำนวนสตรีในแต่ละกลุ่ม	(II) Unadjusted Deviation	(III) ETA	(IV) Adjusted Deviation	(V) BETA
I. อายุของสตรี					
1.1 อายุ 20 ปี หรือต่ำกว่า	5	−2.61		−2.99	
1.2 อายุระหว่าง 21–25 ปี	25	−2.09		−1.75	
1.3 อายุระหว่าง 26–30 ปี	23	−1.00		−0.76	
1.4 อายุระหว่าง 31–35 ปี	12	−0.70		−0.44	
1.5 อายุระหว่าง 36–40 ปี	21	0.29		0.21	
1.6 อายุ 41 ปี ขึ้นไป	46	1.97		1.65	
			.64		.56
II. อาชีพของสามี					
2.1 ทำการกสิกรรม	96	0.41		−0.07	
2.2 ข้าราชการ	20	−1.51		−0.07	
2.3 อื่นๆ	16	−0.55		0.52	
			.28		.08
III. รายได้ของครอบครัวต่อปีต่อสมาชิก 1 คน					
3.1 รายได้ 1,500 บาท หรือต่ำกว่า	28	1.31		0.92	
3.2 ระหว่าง 1,501–2,500 บาท	26	0.04		−0.07	
3.3 ระหว่าง 2,501–4,500 บาท	42	0.08		−0.01	
3.4 4,501 บาท ขึ้นไป	36	−1.14		−0.66	
			.33		.22

$R^2 = .510$ $F = 21.106$

ที่มา: ใช้ข้อมูลจากการสำรวจของ M. Santikarn, "Fertility : A Case Study of Ngao District", SEAPRAP Research Report, 1980.

mean) แต่สัมประสิทธิ์ของตัวแปร Dummy แสดงถึงสัมประสิทธิ์ที่เปรียบเทียบกับกลุ่มที่ได้ละเว้นไป (omitted categories) ก่อนที่จะเปรียบเทียบสัมประสิทธิ์ของ MCA กับ MRA ขอย้อนทบทวนถึงความหมายของสัมประสิทธิ์ของตัวแปร dummy ใน MRA ก่อน

แผนภาพที่ 1 ปัจจัยที่กำหนดภาวะเจริญพันธุ์ของสตรีสมรสแล้ว ในชนบทอำเภองาว จังหวัดลำปาง (เลือกเฉพาะคุณสมบัติบางประการ)



ที่มา : จากตารางที่ 1

สมมติว่าเราต้องการอธิบายจำนวนบุตรเกิดรอด (C) โดยพิจารณาปัจจัย 2 พวก คือ รายได้ของครอบครัว (Y) และขนาดของฟาร์มหรือที่ดิน (L) สมการ regression ซึ่งมีขนาดของฟาร์มหรือที่ดินเป็นตัวแปรประเภท dummy จะมีลักษณะดังนี้

(a) $C_i = k + aY_i + b_{11}L_{1i} + b_{12}L_{2i} + b_{13}L_{3i} + e_i$

โดยที่ k เป็นตัวคงที่

e_i เป็น error of estimate

เนื่องจาก L_{ji} ในฐานะที่เป็นตัวแปร dummy มีคุณสมบัติพิเศษ คือ

$$\sum_{j=1}^{3} L_{ji} = 1$$

ซึ่งจะทำให้เกิด perfect linear multiple correlation ขึ้น วิธีแก้ไขปัญหามีหลายวิธี* เช่น บังคับให้ตัวคงที่ k มีค่าเป็นศูนย์

(b) $C_i = aY_i + b_{21}L_{1i} + b_{22}L_{2i} + b_{23}L_{3i} + e_i$ หรือ

โดยการละทิ้งตัวแปร dummy ไป 1 ตัว

(c) $C_i = k + aY_i + b_{31}L_{1i} + b_{32}L_{1i} + e_i$

โดยเฉพาะสมการ (c) มีผู้นิยมใช้กันมาก เพราะไม่ได้บังคับให้เส้นความสัมพันธ์ให้ผ่านจุดเริ่มต้น

อย่างไรก็ดี ตัวสัมประสิทธิ์ใน 2 สมการนี้ให้ความหมายต่างกัน b_{2I} ในสมการที่ (b) วัดอิทธิพลของขนาดของที่ดินที่มีต่อจำนวนบุตรซึ่งเกิดรอดซึ่งอันที่จริงแล้ว b_{21} ในสมการ (b) ทำหน้าที่แยกสมการที่ (b) ออกเป็น 3 สมการย่อย โดยมี b_{21} เป็น intercept นั่นเอง ส่วน b_{31} นั้นวัดผลกระทบของขนาดของที่ดินที่มีต่อจำนวนบุตร เมื่อเปรียบเทียบกับผลกระทบที่จะเกิดจากขนาดของที่ดินขนาด (c) ซึ่งถูกละจากสมการ (omitted category) นั่นคือ

สมมติว่า L_{1i} ผู้มีที่ดินต่ำกว่า 3 ไร่ เรียกย่อ ๆ ว่าผู้มีที่ดินขนาดเล็ก

L_{2i} ผู้มีที่ดินระหว่าง 3–6 ไร่ หรือผู้มีที่ดินขนาดกลาง

---

* ดู Daniel B. Suits, "Use of Dummy Variables in Regression Equations" *American Statistical Association Journal*, (December, 1951,) p. 549.

L_{3i} ผู้มีที่ดินมากกว่า 6 ไร่ หรือผู้มีที่ดินขนาดใหญ่

$b_{31} = b_{21} - b_{23}$ คือผลต่างของจำนวนบุตรเกิดรอดระหว่างผู้มีที่ดินต่ำกว่า 3 ไร่ กับผู้มีที่ดินมากกว่า 6 ไร่ หรือระหว่างผู้มีที่ดินขนาดเล็กและผู้มีที่ดินขนาดใหญ่ (ซึ่งเกิดจากอิทธิพลของขนาดของฟาร์มหรือที่ดิน)

$b_{32} = b_{22} - b_{23}$ คือผลต่างของจำนวนบุตรระหว่างผู้มีที่ดินขนาดกลางกับผู้มีที่ดินขนาดใหญ่ (ซึ่งเกิดจากอิทธิพลของขนาดของฟาร์มหรือที่ดิน)

ที่นี้ลองมาพิจารณาดูสัมประสิทธิ์ของ MCA

$$C_i = \bar{C} + aY_i + \sum_{j=1}^{3} \beta_i L_{ji} + e_i$$

โดยที่ $\bar{C}$ เป็น grand mean ของจำนวนบุตรเกิดรอด

β_i เป็นสัมประสิทธิ์ MCA ซึ่งวัดการเปลี่ยนแปลงของจำนวนบุตรเกิดรอดจาก grand mean ในกรณีที่มีที่ดินขนาดต่าง ๆ สมมติว่า

$\bar{C} = 3.3$

$\beta_1 = .70$

$\beta_2 = .20$

$\beta_3 = -.30$

เราสามารถคาดคะเนได้ว่า ผู้มีที่ดินขนาดเล็กจะมีบุตร 4 คน (3.3 + .70) ผู้มีที่ดินขนาดกลางจะมีบุตร 3.5 (3.3 + .20) คน ผู้มีที่ดินขนาดใหญ่จะมีบุตรเพียง 3 คน (3.3–.30)

ตัวคงที่ใน MRA มีค่าเท่ากับผลรวมของ grand mean และ B_3 นั่นเอง

โมเดลของ MCA

โมเดลของ MCA มีดังนี้ คือ.–

1. $$Y_i = \bar{Y} + \sum_{j=i}^{r} \alpha_j X_{ji} + \sum_{k=1}^{s} \beta_k Z_{ki} + e_i$$

2. $$\sum_{j=1}^{r} X_{ji} = \sum_{k=1}^{r} Z_{ki} = 1,$$ สำหรับทุกกรณี (observation)

3. $\sum_{j=1}^{r} \alpha_j \bar{X}_j = \sum_{k=1}^{s} \beta_k \bar{Z}_k = 0$

Y_i = ตัวแปรตามของ observation ที่ i

$\bar{Y}$ = Grand mean ของตัวแปรตาม

X_{ji} = ตัวแปรอิสระ X ซึ่งอยู่ในกลุ่ม (category) j ของ observation i

α_j = สัมประสิทธิ์ของกลุ่มที่ j ของตัวแปร X

r = จำนวนกลุ่มของตัวแปร X

Z_{ki} = ตัวแปรอิสระ Z ซึ่งอยู่ในกลุ่ม (category) k ของ observation i

β_k = สัมประสิทธิ์ของกลุ่มที่ k ของตัวแปร Z

s = จำนวนกลุ่มของตัวแปร

e_i = error of estimate

ทั้งนี้แต่ละกลุ่มของตัวแปรทุกตัวจะมีค่าเป็น 0 หรือ 1 เท่านั้น

สมการที่ 1 แสดงให้ถึงอิทธิพลของกลุ่มของตัวแปรอิสระแต่ละตัวในลักษณะของความบ่ายเบน (deviation) จาก grand mean โดยมิได้คำนึงถึง (adjusted for) ผลของกลุ่มของตัวแปรอิสระตัวอื่น

สมการที่ 2 กำหนดเงื่อนไขว่า ตัวแปรทั้ง 2 ชุด (classifications) คือทั้ง X และ Y มีคุณสมบัติว่าผลรวมของแต่ละชุด classification ของแต่ละ observation จะเท่ากับหนึ่ง ตัวอย่างเช่น ผู้ตอบคนหนึ่ง คือ observation หนึ่ง ถ้าคนนั้นเป็นสมาชิกของกลุ่มใดแล้วจะเป็นสมาชิกของกลุ่มอื่นในชุด (classification) เดียวกันไม่ได้ สมมติว่า X_{ji} เป็นเพศของผู้ตอบ ถ้าผู้ตอบเป็นชายก็จะเป็นหญิงอีกไม่ได้*

ส่วน Model ของ MRA ที่มีตัวแปร dummy มีดังนี้ คือ.–

$$Y_i = k + \sum_{j=1}^{r-1} a_j X_{ji} + \sum_{k=1}^{s-1} b_s Z_{ki} + e_i$$

---

* ให้ตัวแปร x_{1i} มีค่าเป็นหนึ่งถ้าผู้ตอบเป็นเพศชาย และ x_{2i} มีค่าเป็นหนึ่งถ้าผู้ตอบเป็นหญิง ในกรณีที่ผู้ตอบคนหนึ่งเป็นเพศชาย $x_{11} = 1$, $x_{21} = 0$ ดังนั้น $\sum_{j=1}^{2} X_{ji} = 1 + 0 = 1$

ดังนั้นนอกจากเหตุผลที่ว่า MCA สามารถให้คำตอบที่แสดงถึงความบ่ายเบนจากค่ากลางในขณะที่ MRA สามารถให้คำตอบที่เปรียบเทียบกับสัมประสิทธิ์ของตัวที่ถูกละเว้น MCA ยังมีประโยชน์กว่า MRA ในข้อที่ว่า MCA ให้คำตอบเพียงอันเดียว (unique solution) ในขณะที่ MRA ที่ใช้ตัวแปร dummy จะมี r คูณ s วิธีที่จะเสนอคำตอบ

ความคล้ายคลึงของทั้งสองโมเดลทำให้สามารถเปลี่ยนสัมประสิทธิ์ของโมเดลหนึ่งไปหาสัมประสิทธิ์ของอีกโมเดลหนึ่ง สามารถพิสูจน์ได้ว่า (ดูการพิสูจน์ในภาคผนวก)

$a_j = \alpha_j - \alpha_r$ โดยที่ a_j, b_k เป็นสัมประสิทธิ์ MCA

$b_k = \beta_k - \beta_s$ α_j, β_k เป็นสัมประสิทธิ์ MCA

หรือ $\alpha_j = a_j + \alpha_r$

$\beta_k = b_k + \beta_s$

โดยที่ r, s เป็นกลุ่มที่ถูกละเว้นในการทำ MRA ที่ใช้ตัวแปร dummy แต่เนื่องจาก MRA ไม่ให้ค่าสัมประสิทธิ์ของกลุ่มที่ถูกละเว้น ทำให้เราไม่สามารถหาค่า α_r, β_s ได้จากสัมประสิทธิ์dummy โดยตรง แต่มีวิธีหา α_r, β_s ดังนี้คือ.–

จากสมการที่ 3 ของโมเดล MCA

$$\sum_{j=1}^{r} \alpha_j \bar{X}_j = \sum_{j=1}^{r-1} \alpha_j \bar{X}_j + \alpha_r \bar{X}_r = 0$$

$$= \sum_{j=1}^{r-1} \alpha_j \bar{X}_j + \alpha_r \left(1 - \sum_{j=1}^{r-1} \bar{X}_j\right) = 0$$

$$= \alpha_r + \sum_{j=1}^{r-1} (\alpha_j - \alpha_r) \bar{X}_j = 0$$

$$\alpha_r = -\sum_{j=1}^{r-1} (\alpha_j - \alpha_r) \bar{X}_j = -\sum_{j=1}^{r-1} a_j \bar{X}_j$$

ในทำนองเดียวกัน

$$\beta_s = -\sum_{k=1}^{s-1} (\beta_k - \beta_s)\,\bar{Z}_k = -\sum_{k=1}^{s-1} b_k\,\bar{Z}_k$$

ดังนั้น ถ้าเราทราบค่าของสัมประสิทธิ์ MRA (a_j, b_k) เราก็จะสามารถรู้ค่าของสัมประสิทธิ์ MCA (α_j, β_k) ได้ดังนี้คือ

$$\alpha_j = a_j - \sum_{j=1}^{r-1} a_j\,\bar{X}_j$$

$$\beta_k = b_k - \sum_{k=1}^{s-1} b_k\,\bar{X}_k$$

ตัวอย่างที่ 1 โดยการใช้ข้อมูลเดียวกับตารางที่ 1 และเลือกมาเฉพาะตัวแปรอายุ ที่มีตัวแปร dummy มาเปลี่ยนเป็นสัมประสิทธิ์ MCA จะได้ผลลัพธ์ดังปรากฏในตารางที่ 2

ดังนั้น $\alpha_1 = -4.635 - [\,(.037)(-4.63) + (.189)(-3.399)$

$(.174)(-2.407) + (.090)(-2.093) + (.159)(-1.440)\,]$

$= -4.635 - [-1.641]$ โดยที่ α_r หรือในกรณีนี้คือ $\alpha_6 = 1.649$

$\alpha_1 = 2.986$ ซึ่งก็คือสัมประสิทธิ์ MCA ของสตรีอายุ 20 ปี หรือต่ำกว่า (ตาราง 2, สดมภ์ 4) α_2, α_3, α_4, α_5 ก็จะหาได้โดยเอา α_6 ไปบวกกับ α_2, α_3, α_4, α_5 ในสดมภ์ที่ 2 ซึ่งจะทำให้ได้ตัวเลขในสดมภ์สุดท้ายของตารางที่ 2

ในการหาค่า eta^2 เราจะต้องรู้ค่าบ่ายเบนจากค่ากลาง (unadjusted deviation from grand mean) ซึ่งหาได้โดยการลบ grand mean ($\bar{X}$) ออกจาก group mean ($\bar{X}_j$) หรือหาได้โดยเอาการรวมผลคูณของสดมภ์ที่ 2 และ ที่ 3^2 (ยกกำลังสอง ตารางที่ 2) แล้วหารด้วย variance ของตัวแปรตามซึ่งในที่นี้คือ variance ของจำนวนครั้งที่ตั้งครรภ์ ซึ่งเท่ากับ 6.44 ส่วน beta^2 หาได้โดยการรวมผลคูณของสดมภ์ที่ 2 ในตารางที่ 2 กับกำลังสองของสดมภ์ที่ 4 แล้วหารด้วย variance ของจำนวนครั้งที่ตั้งครรภ์ ถอดรากที่ 2 จะได้ eta = .64 และ beta = .56 ดังในตารางที่ 1

ตารางที่ 2 อิทธิพลของอายุสตรี (จำนวนครั้งที่ตั้งครรภ์) Grand mean = 3.61

ตัวแปรอิสระ	สดมภ์ I สัมประสิทธิของตัวแปร(dummy) (a_j)	สดมภ์ II การกระจายของ observation ในแต่ละกลุ่มอายุ ($\bar{X}_j$)	สดมภ์ III Unadjusted deviation ($\bar{X}_j - \bar{X}$)	สดมภ์ IV Adjusted deviation (a_j)
อายุสตรี				
20 ปีหรือต่ำกว่า	−4.635	.037	−2.61	−2.986
อายุระหว่าง 21–25 ปี	−3.399	.189	−2.09	−1.750
อายุระหว่าง 26–30 ปี	−2.407	.174	−1.00	−0.758
อายุระหว่าง 31–35 ปี	−2.039	.090	−0.70	.444
อายุระหว่าง 36–40 ปี	−1.440	.159	0.29	.209
อายุ 41 ปีขึ้นไป (omitted categary)		.348	1.97	1.649

ที่มา : เช่นเดียวกับตารางที่ 1

ในเมื่อ MRA และ MCA มีความสัมพันธ์กันอย่างใกล้ชิด ผู้วิจัยสามารถเลือกใช้ได้ตามความเหมาะสมของข้อมูล ในกรณีที่ผู้วิจัยต้องการเปรียบเทียบกับอิทธิพลของตัวแปรที่มีเพียง 2 กลุ่ม เช่น เพศชายกับเพศหญิง หมู่บ้านที่มีไฟฟ้ากับหมู่บ้านที่ไม่มีไฟฟ้า MRA ที่ใช้ตัวแปร dummy จะมีประสิทธิภาพกว่า เพราะให้ผลเปรียบเทียบทันที ในกรณีที่ตัวแปรมีหลายกลุ่ม และต้องการรู้ผลกระทบของกลุ่มทุกกลุ่ม MCA จะให้ภาพที่ชัดเจนกว่า ในกรณีที่ต้องการสัมประสิทธิ์ MCA แต่ไม่มี program สำเร็จรูปให้ใช้ เราก็สามารถอาศัยการเปลี่ยนสัมประสิทธิ์ MRA เป็น MCA ได้โดยง่ายดังวิธีที่กล่าวมาแล้ว

ภาคผนวก : การพิสูจน์ความสัมพันธ์ระหว่างสัมประสิทธิ์ MCA กับ MRA ที่มีตัวแปรอิสระเป็นตัวแปร dummy

1. โมเดลของ MCA

$$(1)\quad Y_i = \bar{Y} + \sum_{j=1}^{r} \alpha_j X_{ji} + \sum_{k=1}^{s} \beta_k Z_{ki} + e_i$$

$$(2)\quad \sum_{j=1}^{r} X_{ji} = \sum_{k=1}^{s} Z_{ki} = 1 \; ; \forall_i$$

$$(3)\quad \sum_{j=1}^{r} \alpha_j \bar{X}_j = \sum_{k=1}^{s} \beta_k \bar{Z}_k = 0$$

สูตรสมการที่ 1 ด้วยตัวแปรทุกกลุ่ม คือ $\sum X_{ji}$ และ $\sum Z_{ki}$ แต่เนื่องจากทุก observation มีค่าเป็น 0 หรือ 1 ดังนั้น $\sum_{i=1}^{n} X_{ji}^2 = \sum_{i=1}^{n}$ และถ้าสมการที่ (2) เป็นจริง พร้อมทั้ง $\sum_{j=1}^{n} X_{ji} X_{mi} = 0,\ j \neq m$ แล้ว เราจะได้ชุดของสมการ (1) ดังต่อไปนี้คือ

ชุดของ X_j

$$\begin{cases} \sum X_{1i} Y_i = \bar{Y} \sum X_{1i} + \alpha_1 \sum X_{1i} + \sum_{k=1}^{s} \beta_k \left(\sum X_{1i} Z_{ki}\right)^* \\ \vdots \\ \sum X_{ji} Y_i = \bar{Y} \sum X_{ji} + \alpha_j \sum X_{ji} + \sum_{k=1}^{s} \beta_k \left(\sum X_{ji} Z_{ki}\right) \\ \vdots \\ \sum X_{ri} Y_i = \bar{Y} \sum X_{ri} + \alpha_r \sum X_{ri} + \sum_{k=1}^{s} \beta_k \left(\sum X_{ri} Z_{ki}\right) \end{cases}$$

---

* ยกตัวอย่างเช่น

$$\Sigma X_{1i} Y_i = \bar{Y} \Sigma X_{1i} + \alpha_1 \Sigma X_{1i} \Sigma X_{1i} + \alpha_2 \Sigma X_{1i} \Sigma X_{2j} + \ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \alpha_r \Sigma X_{1i} \Sigma X_{ri} + \sum_{k=1}^{s} \beta_k \left(\Sigma X_{1i} Z_{ki}\right)$$

เทอมที่ 2 ของทางขวามือ $\alpha_1 \Sigma X_{ij}^2 = \alpha_1 \Sigma X_{ij}$

เทอมที่ 3 จนกระทั่งเทอมก่อนเทอมสุดท้ายจะกลายเป็นศูนย์ไปหมด ดังนั้น

$$\Sigma X_{1i} Y_i = \bar{Y} \Sigma X_{1i} + \alpha_1 \Sigma X_{1i} + \sum_{k=1}^{s} \beta_k \left(\Sigma X_{1i} Z_{ki}\right)$$

ชุดของ Z_k

$$\begin{cases} \sum Z_{1i}\, Y_i = \bar{Y} \sum Z_{1i} + \sum_{j=1}^{r} \alpha_j \left(\sum Z_{ji}\, X_{ji}\right) + \beta_1 \sum Z_{ji} \\ \vdots \\ \sum Z_{ki}\, Y_i = \bar{Y} \sum Z_{ki} + \sum_{j=1}^{r} \alpha_j \left(\sum Z_{ki}\, X_{ji}\right) + \beta_k \sum Z_{ki} \\ \vdots \\ \sum Z_{si}\, Y_i = \bar{Y} \sum Z_{si} + \sum_{j=1}^{r} \alpha_j \left(\sum Z_{si}\, X_{ji}\right) + \beta_s \sum Z_{si} \end{cases}$$

แต่เนื่องจาก (4) $\sum X_{ji} = n_j$ ผลรวมของค่าตัวแปร ในกลุ่มใดกลุ่มหนึ่งก็คือ จำนวนคนในกลุ่มนั้น* และ

(5) $\sum X_{ji}\, Z_{ki} = n_{jk}$ ผลรวมของผลคูณของ dummy variables 2 กลุ่ม (ที่ไม่ใช่ Classification เดียวกัน) ก็คือ จำนวนคนที่มีคุณสมบัติทั้ง 2 ประเภท ซึ่งก็คือ จำนวนคนที่เป็นสมาชิกของทั้ง 2 กลุ่ม ดังนั้น

$$(6)\quad \sum_{i=1}^{n} X_{ji}\, Y_i = \sum_{i=1}^{n_{j.}} Y_{ji} = n_{i.}\, \bar{Y}_i$$

โดยแทนค่าสมการที่ 6 แล้วย้ายเทอมที่ 1 ทางขวามือไปข้างซ้ายมือ normal equation ชุดใหม่ก็จะมีลักษณะดังนี้

ชุดของ X_j

$$\begin{cases} n_{1.}\ (\bar{Y}_{1.} - \bar{Y}) = n_{1.}\ \alpha_1 + \sum_{k=1}^{s} n_{1k}\ \beta_k \\ \vdots \\ n_{j.}\ (\bar{Y}_{j.} - \bar{Y}) = n_{j.}\ \alpha_j + \sum_{k=1}^{s} n_{jk}\ \beta_k \\ \vdots \\ n_{r.}\ (\bar{Y}_{r.} - \bar{Y}) = n_{r.}\ \alpha_r + \sum_{k=1}^{s} n_{nk}\ \beta_k \end{cases}$$

---

* $n_{j.}$ = จำนวนสมาชิกในกลุ่มที่ j ของตัวแปร X

$n_{.k}$ = จำนวนสมาชิกในกลุ่มที่ k ของตัวแปร Z

ชุดของ Z_k

$$
\begin{aligned}
n_{.1}\ (\bar{Y}_{.1}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_j\, n_{j1} + n_{.1}\,\beta_1 \\
&\ \ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
n_{.k}\ (\bar{Y}_{.k}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_j\, n_{jk} + n_{.k}\,\beta_k \\
&\ \ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
n_{.s}\ (\bar{Y}_{.s}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_j\, n_{js} + n_{.s}\,\beta_s
\end{aligned}
$$

อย่างไรก็ดี ระบบของสมการข้างต้นนี้ไม่สามารถให้คำตอบได้ เพราะมีลักษณะlinearly dependent ในแต่ละชุด ดังนั้นจึงจำเป็นต้องทิ้งสมการในแต่ละชุดไปหนึ่งสมการแล้วแทนที่ด้วยสมการที่ 3 จะได้ normal equations ชุดใหม่อีกดังนี้

ชุดของ X_j

$$
\begin{aligned}
n_{1.}\ (\bar{Y}_{1.}-\bar{Y}) &= n_{1.}\ \alpha_1 + \sum_{k=1}^{s} n_{1k}\,\beta_k \\
&\ \ \vdots \\
n_{j.}\ (\bar{Y}_{j.}-\bar{Y}) &= n_{j.}\ \alpha_j + \sum_{k=1}^{s} n_{jk}\,\beta_k \\
&\ \ \vdots \\
n_{r-1.}\ (\bar{Y}_{r-1.}-\bar{Y}) &= n_{r-1.}\ \alpha_{r-1} + \sum_{k=1}^{s} n_{(r-1)k}\,\beta_k \\
0 &= \sum_{j-1}^{r} n_{j.}\ \alpha_j
\end{aligned}
$$

ชุดของ Z_k

$$
\begin{aligned}
n_{.1}\ (\bar{Y}_{.1}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_j\, n_{j1} + n_{.1}\,\beta_1 \\
&\ \ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
n_{.k}\ (\bar{Y}_{.k}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_j\, n_{jk} + n_{.k}\,\beta_k \\
&\ \ \vdots \qquad\qquad \vdots \\
n_{.s-1}\ (\bar{Y}_{.s-1}-\bar{Y}) &= \sum_{j=1}^{r} \alpha_i\, n_{.i(s-1)} + n_{.s-1}\,\beta_{s-1} \\
0 &= \sum_{k=1}^{s} n_{.k}\ B_k
\end{aligned}
$$

โดยการหาค่าจากสมการข้างต้นนี้ ก็จะได้สัมประสิทธิ์ α_i, β_k ของ MCA โดยตรง

ส่วนโมเดล MRA ที่ใช้ตัวแปร dummy นั้นแตกต่างจากโมเดล MCA ตรงที่ปัญหา linear dependence อันเกิดจากเงื่อนไขที่ว่า $\sum_{j=1}^{r} X_{ji} = \sum_{k=1}^{s} Z_{k1} = 1$ นั้นได้ถูกกำจัดไปโดยการละทิ้งตัวแปร dummy ของตัวแปรแต่ละชุด (classification) ไป 1 ตัวตั้งแต่ต้น กล่าวคือ

$$Y_j = k + \sum_{j=1}^{r-1} a_j X_{ji} + \sum_{k=1}^{s-1} b_k Z_{ki} + e_i$$

X_r และ Z_s เป็นกลุ่มที่ถูกละเว้น (omitted category)

k เป็นค่าที่รวมเอาสัมประสิทธิ์ของกลุ่มสุดท้ายทั้ง 2 กลุ่ม ของตัวแปรทั้ง 2 ชุดไว้

ดังนั้น normal equations ของ MRA มีดังนี้คือ

$$\sum Y_i = kN + \sum_{j=1}^{r-1} a_j \left(\sum X_{ji}\right) + \sum_{k=1}^{s-1} b_k \left(\sum Z_{ki}\right)$$

$$\sum Y_i X_{1i} = k \sum X_{1i} + a_j \sum X_{1i} + \sum_{k=1}^{s-1} b_k \left(\sum X_{1i} Z_{ki}\right)$$

$$\sum Y_i X_{ji} = k \sum X_{ji} + a_j \sum X_{ji} + \sum_{k=1}^{s-1} b_k \left(\sum X_{ji} Z_{ki}\right)$$

$$\sum Y_i X_{(r-1)i} = K \sum X_{(r-1)i} + a_{r-1} \sum X_{(r-1)i} + \sum_{k=1}^{s-1} \left(\sum X_{(r-1)} Z_{ki}\right)$$

$$\sum Y_i Z_{1i} = K \sum Z_{1i} + \sum_{j=1}^{r-1} a_j \left(\sum X_{ji} Z_{1i}\right) + b_1 \sum Z_{1i}$$

$$\sum Y_i Z_{ki} = k \sum Z_{ki} + \sum_{j=1}^{r-1} a_j \left(\sum X_{ji} Z_{ki}\right) + b_k \sum Z_{ki}$$

$$\sum Y_i Z_{(s-1)i} = k \sum Z_{(s-1)i} + \sum_{j=1}^{s-1} a_j \sum X_{ji} Z_{(s-1)i} + b_{s-1} \sum Z_{(s-1)i}$$

พยายามจัด normal equations ของตัวแปร dummy ให้มีลักษณะเหมือนของ MCA โดยแทนค่า k

$$(7)\quad k = \bar{Y} - \sum_{j=1}^{r-1} a_j \bar{X}_j = \sum_{k=1}^{s-1} b_k \bar{Z}_k$$

ลงไปใน general equation ของ X_j ของ MRA

$$(8)\quad \sum Y_i X_{ji} = n_{j.} \bar{Y} - n_{j.} \left(\sum_{j=1}^{r-1} a_j \bar{X}_j\right) - n_{j.} \left(\sum_{k=1}^{s-1} b_k \bar{Z}_k\right) + n_{j.} a_{j.} + \sum_{k=1}^{s-1} n_{kj} b_k$$

แต่ $(9)\quad \sum Y_j X_{ji} = n_{j.} \bar{Y}_{j.}$

แทนค่า (9) ใน (8) แล้วลบด้วย $n_{j.} \bar{Y}$ ทั้งสองข้างจะได้

$$(10)\quad n_{j.} (\bar{Y}_{j.} - \bar{Y}) = - n_{j.} \left(\sum_{j=1}^{r-1} a_j \bar{X}_j\right) - n_{j.} \left(\sum_{k=1}^{s-1} b_k \bar{Z}_K\right) + n_{j.} a_j + \sum_{k=1}^{s-1} n_{jk} b_k$$

ลองย้อนกลับไปเปรียบเทียบสมการที่ 10 กับ สมการในเซ็ท X_j ของ MCA ซึ่งเท่ากับ

$$(11)\quad n_{j.} (\bar{Y}_{j.} - \bar{Y}) = n_{j.} \alpha_j + \sum_{k=1}^{s-1} n_{jk} \beta_k + n_{js} \beta_s$$

บวกและลบด้วย $n_{j.} \alpha_r$ และ $\sum_{k=1}^{s-1} n_{jk} \beta_s$ เข้าในทางข้างขวาของสมการ 11 จะได้

$$(12)\quad n_{j.} (\bar{Y}_{j.} - \bar{Y}) = n_{j.} \alpha_r + \sum_{k=1}^{s-1} n_{jk} \beta_s + n_{j.} (\alpha_j - \alpha_r) + n_{js} \beta_s + \sum_{k=1}^{s-1} n_{jk} (\beta_k \beta_s)$$

แต่ (13) $\sum_{k=1}^{s-1} n_{jk}\,\beta_{jk}\beta_s + n_{js}\,\beta_s = \sum_{k=1}^{s} n_{jk}\,\beta_s = \beta_s \sum_{k=1}^{s} n_{jk} = n_{j.}\,\beta_s$

แทนค่าลงใน (12) จะได้

$$(14)\quad n_{j.}\,(\bar{Y}_{j.} - \bar{Y}) = n_{j.}\,\alpha_r + n_{j.}\,\beta_s + n_{j.}\,(\alpha_j - \alpha_r) + \sum_{k=1}^{s-1} n_{jk}\,(\beta_k - \beta_s)$$

จากสมการที่ 3 เรารู้ว่า

$$(15a)\quad \alpha_r = -\sum_{j=1}^{r-1} (\alpha_j - \alpha_r)\,\bar{X}_j$$

$$(15)\quad \beta_s = -\sum_{k=1}^{s-1} (\beta_k - \beta_s)\,\bar{Z}_k$$

แทนค่า (15a) และ (15b) เข้าไปในสมการ (14) จะได้

$$(16)\quad n_{j.}\,(\bar{Y}_{j.} - \bar{Y}) = -\,n_{j.}\left(\sum_{i=1}^{r-1} (\alpha_j - \alpha_r)\,\bar{X}_j\right) - n_{j.}\left(\sum_{k=1}^{s-1} (\rho_k - \beta_s)\,\bar{Z}_k\right) + n_{j.}\,(\alpha_j - \alpha_r) + \sum_{k=1}^{s-1} n_{jk}\,(\beta_k - \beta_s)$$

จะเห็นได้ว่า สมการ (16) ของ MCA มีลักษณะเกือบเหมือนสมการ (10) ของ MRA ทุกประการ นอกจากว่าในสมการที่ 16 $\alpha_j - \alpha_r$ แทนที่ a_j และ $\beta_k - \beta_s$ แทนที่ b_k ทั้งคู่จะมี $r + s - 2$ สมการ ซึ่งจะให้คำตอบเพียงคำตอบเดียว ดังนั้น

จึงสรุปได้ว่า $a_j = \alpha_j - \alpha_r$

$b_k = \beta_k - \beta_s$ นั่นเอง

บทความและหนังสืออ้างอิง

Andrews, Frank M., James N. Morgan, John A. Sonquist and Laura Klem. *Multiple Classification Analysis* (2nd edition). Institute for Social Research, University of Michigan, 1973.

ESCAP. "Multiple Classification Analysis and Its Application to the 1974 Fiji Fertility Survey" in *Report and Selected Papers for Regional Workshop on Techniques of Analysis of World Fertility Survey Data*, Asian Population Studies Series No. 44. Bangkok, 1979.

Gordon, R.A. "Issues .in Multiple Regression". *American Journal of Sociology*, Vol. 3, No. 5, (1968), pp. 592–601,

Hodge, R.W. "Multiple Classification and Dummy Variables : An Example" a lecture given at the Multivariate Data Analysis Workshop, 13 August, 1979.

Santikarn, Mingsarn. "Fertility and Family Planning : A Case Study of Ngao District" SEAPRAP report, 1980.

Suits, Daniel B. "Use of Dummy Variables in Regression Equations" *American Statistical Association Journal*, Vol. 57, (December, 1957), pp. 548–551,