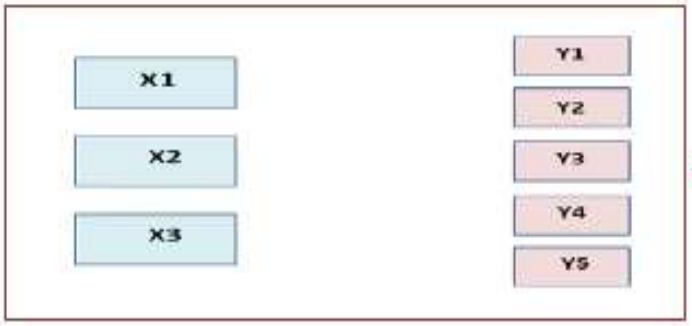
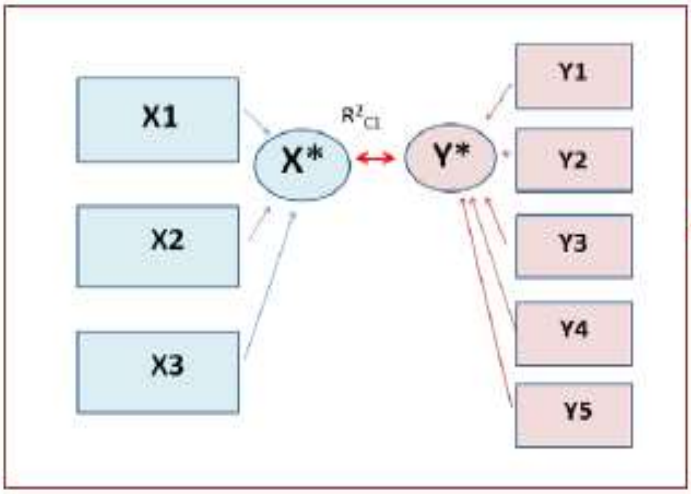
27

Canonical Correlation Analysis (Cancorr)

Canonical correlation analysis เป็นเทคนิคทางสถิติที่ใช้ในการหาความสัมพันธ์ระหว่างตัวแปรสองกลุ่ม กลุ่มแรกจะเป็นกลุ่มของตัวแปรอิสระ (สมมติประกอบด้วย X1,X2, X3) อีกกลุ่มหนึ่งจะเป็นกลุ่มของตัวแปรตาม (สมมติประกอบด้วย Y1 , Y2 , Y3, Y4 , Y5) ซึ่งในแต่ละกลุ่มอาจจะประกอบด้วยจำนวนตัวแปรไม่เท่ากัน ในการวิเคราะห์ cancorr นักสถิติอาจไม่จำเป็นต้องกำหนดว่ากลุ่มตัวแปรใดเป็นกลุ่มของตัวแปรอิสระ / กลุ่มของตัวแปรตามก็ได้ (Tabachnick & Fidell , 2007)กลไกของ cancorr จะอาศัยการสร้างตัวแปรสังเคราะห์ (synthetic variable /latent variable) หรือที่เรียกว่า variate ซึ่งเป็น linear combination ของตัวแปรในแต่ละกลุ่ม (ตัวแปรสังเคราะห์ของกลุ่มตัวแปรอิสระเรียกว่า covariate canonical variable ส่วนตัวแปรสังเคราะห์ของกลุ่มตัวแปรตามเรียกว่า dependent canonical variable) โดย covariate canonical variable และ dependent canonical variable ที่สร้างขึ้นจะมีความสัมพันธ์เชิงเส้นตรงสูงสุด หากมี variance ที่ยังไม่สามารถอธิบายได้ด้วย variate คู่แรก cancorr จะสร้างตัวแปรสังเคราะห์หรือ variate คู่ต่อไปซึ่งเป็น linear combination ของตัวแปรในแต่ละกลุ่มขึ้นมา โดย linear combination ของกลุ่มตัวแปรอิสระจะ orthogonal (มีความเป็นอิสระ / ไม่ขึ้นกัน) กับ linear combination ของกลุ่มตัวแปรตาม โดยที่ความสัมพันธ์เชิงเส้นตรงระหว่าง variate ทั้งสองจะมีค่าสูงสุด หากมี variance ที่ยังไม่สามารถอธิบายได้อยู่อีก ก็จะมีการหาตัวแปรสังเคราะห์หรือ variate คู่ต่อไป ในทางปฏิบัติ cancorr จะหาตัวแปรสังเคราะห์หรือ variate ไม่เกินจำนวนตัวแปรในกลุ่มที่มีจำนวนสมาชิกน้อยกว่า (หากมีตัวแปรอิสระ X1 , X2 , X3 และตัวแปรตาม Y1, Y2, Y3 , Y4 , Y5 ตามตัวอย่างข้างต้น จะมี canonical variable 3 คู่) เพื่อให้เข้าใจได้ง่ายขึ้นเราลองพิจารณาภาพดังต่อไปนี้



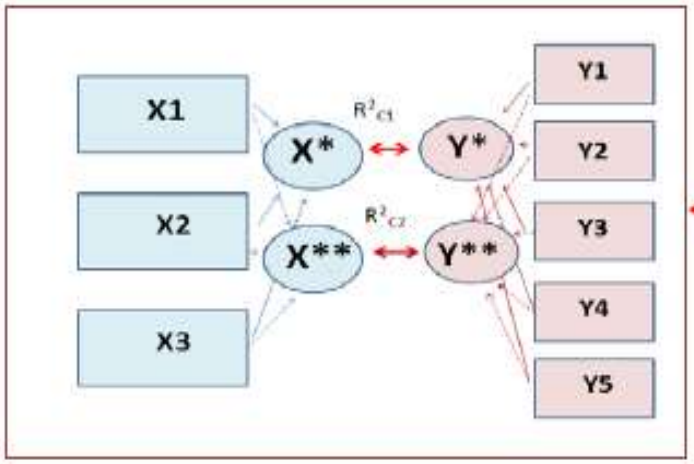
ภาพที่ 1 ด้านซ้ายแสดงกลุ่มของตัวแปรสองกลุ่ม คือกลุ่มของตัวแปร X's (มีจำนวนตัวแปร 3 ตัว) และกลุ่มของตัวแปร Y's (มีจำนวนตัวแปร 5 ตัว)



Canonical correlation analysisจะหา variate/synthetic variable(X*)ที่เป็นตัวแทนของ X's และ(Y*)ที่เป็นตัวแทนของ Y's ที่ซึ่งความสัมพันธ์ระหว่าง X* และY*สูงสุด

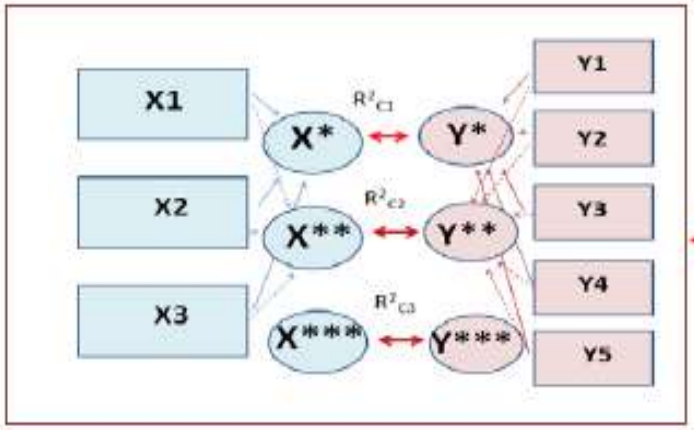
หมายเหตุ:-X*เป็น linear combination ของ X's

Y*เป็น linear combination ของ Y's



เนื่องจากโดยปกติ X* ไม่สามารถอธิบาย variance ของ Y* ได้หมด (R^2_{C1}<1.0) เมื่อเป็นเช่นนี้จะต้องอาศัย variate/ synthetic variable คู่ใหม่ที่ไม่มีความสัมพันธ์กับ variate/synthetic variable X*,Y* เลย

Variate/synthetic variable คู่ใหม่คือ X**,Y**



เนื่องจากอาจมี variance ของ Y's ที่ไม่สามารถอธิบายได้ด้วย X's ที่เหลืออยู่ แม้ variance ส่วนใหญ่จะอธิบายได้ด้วยคู่ X*,Y* และ X**,Y** เมื่อเป็นเช่นนี้ canonical analysis จะต้องหา variate/synthetic variableคู่ต่อไป (ให้ชื่อว่า X***, Y***)

หมายเหตุ :-จำนวน variate จะมีจำนวนสูงสุดเท่ากับจำนวนตัวแปรในกลุ่มที่มีสมาชิกน้อยกว่า ตัวอย่างเช่น ตามภาพ จะมีแค่ variate สามคู่เท่านั้น

หมายเหตุ:- R^2_C คือ Pearson correlation coefficientระหว่าง variate แต่ละคู่ยกกำลังสองนั่นเอง

ต่อไปนี้เป็นตัวอย่างลักษณะปัญหาทางสถิติที่สามารถใช้ cancorr ในการไขปัญหา

ในวงการสุขภาพ : หาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่เกี่ยวข้องกับการออกกำลังกาย (อัตราการไต่ก้าวขึ้นบันได , ความเร็วในการวิ่ง , น้ำหนักที่ยกได้ , จำนวน push-ups ที่ทำได้ต่อนาที) กับกลุ่มของตัวแปรด้านสุขภาพ (ความดันโลหิต , ระดับคลอเรสเตอรอล , ระดับน้ำตาลในเลือด , ดัชนีมวลกาย)

ในวงการธนาคาร : หาความสัมพันธ์ระหว่างจำนวนบัตรเครดิตที่ครอบครัวถือ กับยอดเฉลี่ยการใช้บัตรเครดิตของครอบครัวต่อเดือน , ขนาดของครอบครัว , รายได้รวมของครอบครัว

ในแวดวงสุขภาพจิต : หาความสัมพันธ์ระหว่างมาตรวัดสุขภาพจิต (ความซึมเศร้า ความเครียด ความเหงา) กับการสนับสนุนทางสังคม (โดยรวม หลักๆที่สำคัญ ครอบครัวและเพื่อน)

ในวงการศึกษา : ศึกษาความสัมพันธ์ระหว่างปัจจัยด้านจิตวิทยา (ความเครียด ความวิตกกังวล การเห็นคุณค่าในตัวเอง) กับมาตรวัดความสามารถทางวิชาการ

ในแวดวงการศึกษา : ศึกษาความสัมพันธ์ระหว่างกลยุทธ์ในการเรียนของนักเรียน (อุปนิสัยในการเรียน เทคนิคในการจดโน้ต) ความสามารถเชิงวิชาการในสาขาวิชาต่าง ๆ

ในแวดวงสุขภาพ : ศึกษาความสัมพันธ์ระหว่างปัจจัยด้านการใช้ชีวิต (อาหาร การออกกำลังกาย การนอนหลับ) กับผลต่อสุขภาพ (ความดันโลหิต ระดับคลอเรสเตอรอล ดัชนีมวลกาย)

ในแวดวงการตลาด : ศึกษาความสัมพันธ์ระหว่างข้อมูลด้านประชากรของลูกค้า (อายุ รายได้ ถิ่นที่อยู่) กับความชอบที่มีต่อลักษณะของผลิตภัณฑ์ต่าง ๆ

ในแวดวงสิ่งแวดล้อม : ศึกษาความสัมพันธ์ระหว่างปัจจัยด้านสิ่งแวดล้อม (อุณหภูมิ ระดับมลภาวะ) กับความหลากหลายของพืชและสัตว์ในระบบนิเวศน์

ในแวดวงจิตวิทยา : ศึกษาความสัมพันธ์ระหว่างบุคลิกภาพส่วนบุคคล (การเป็นคนเปิดตัว , ความไม่มั่นคงทางอารมณ์) กับมาตรวัดการทำงาน

วงการประกันภัย : ศึกษาความสัมพันธ์ระหว่างประเภทกรมธรรม์ประกันภัยที่ทำ (ประกันสุขภาพ ประกันชีวิต ประกันอื่น ๆ) กับ ลักษณะของบุคคลที่ทำประกัน (อายุ รายได้ ภูมิหลังด้านการแพทย์ เพศ)

วงการธนาคาร : ศึกษาความสัมพันธ์ระหว่างลักษณะของเครดิตการ์ดกับประเภทบัญชีที่มี เช่น ออมทรัพย์ กระแสรายวัน ฝากประจำ

ในวงการสิ่งแวดล้อม : ศึกษาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่ใช้วัดสุขอนามัยด้านสิ่งแวดล้อม (จำนวนพันธ์สัตว์ที่นับได้ , ความหลากหลายของพันธ์สัตว์ที่พบ , มวลของสารอินทรีย์ , ผลิตภาพของธรรมชาติสิ่งแวดล้อม) กับกลุ่มของตัวแปรที่ใช้วัดสารพิษที่มีในสิ่งแวดล้อม (ความเข้มข้นของธาตุโลหะหนัก ยาฆ่าแมลง สารไดออกซิน)

ในวงการกีฬา : ศึกษาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่ใช้วัดสรีรศาสตร์ (น้ำหนักเป็นปอนด์ , เส้นรอบเอวเป็นนิ้ว , อัตราการเต้นของหัวใจต่อนาที) กับกลุ่มของตัวแปรที่ใช้วัดการออกกำลังกาย (จำนวน chin-up ที่ทำได้ จำนวน sit-up ที่ทำได้ , จำนวน jumping jack ที่ทำได้)

ในวงการจิตเวช : หาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่แสดงแรงสนับสนุนทางสังคม (สังคมโดยรวม , ครอบครัว , เพื่อน , ที่สำคัญอื่น ๆ) กับกลุ่มของตัวแปรที่แสดงสุขภาพจิต (ความซึมเศร้า , ความว้าเหว่ , ความเครียด)

ในวงการบริหารทรัพยากรมนุษย์ : หาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่แสดงลักษณะงาน (ความหลากหลายของงาน , Feedback , ความเป็นอิสระในการทำงาน) กับกลุ่มของตัวแปรที่แสดงความพึงพอใจในงาน (ความพึงพอใจในเส้นทางการทำงาน , ความพึงพอใจในฝ่ายจัดการและหัวหน้างาน , ความพึงพอใจในรายได้และสวัสดิการ

ในวงการจิตเวช : ศึกษาความสัมพันธ์ระหว่างตัวแปรที่ใช้ทดสอบบุคลิกภาพในมิติที่เป็น MMPI-The Minnesota Multiphasic Inventory (Hypochondriasis -คิดว่าตัวเองป่วย , Depression -ความซึมเศร้า , Hysteria-ขาดความรักไม่ได้ , Psychopathic Deviate -ต่อต้านสังคม ฯ) กับกลุ่มของตัวแปรที่ใช้ทดสอบบุคลิกภาพในมิติที่เป็น NEO (Neuroticism-การเตรียมความพร้อม , Extraversion-อารมณ์ดีเป็นมิตร , Openness-เปิดรับสิ่งใหม่ๆ)

สิ่งที่ต้องทำความเข้าใจเกี่ยวกับการวิเคราะห์ cancorr

1. ผลที่ได้จากการวิเคราะห์จะเป็น variate หลายคู่ คู่แรกจะเป็นคู่ที่มี canonical correlation (R_C) สูงสุด คู่ต่อ ๆไปจะมี canonical correlation ที่มีขนาดเล็กลดหลั่นกันลงไป อย่างไรก็ตาม นักสถิติต้องทำใจว่า ค่า R_C ที่ได้อาจมีค่าไม่สูงมาก เกณฑ์ที่อยู่ในระดับที่น่าพอใจได้แก่กรณี R_C มีค่าตั้งแต่ 0.30 ขึ้นไป (variate ในกลุ่มแรกสามารถอธิบายความผันผวนของ variate ในกลุ่มที่สองได้ร้อยละเก้า หรือ 0.3^2)
2. ความสัมพันธ์ระหว่าง variate แต่ละคู่ก็คือความสัมพันธ์ในมิติที่แตกต่างกัน ดังนั้นความสัมพันธ์ของ Variate ในคู่แรกจะอยู่ในมิติที่แตกต่างจากความสัมพันธ์ของ variate ของคู่อื่น ๆ
3. แม้จะมี variate/ $R^2{}_C$ หลายคู่ แต่ละคู่แสดงมิติหนึ่งของความสัมพันธ์ระหว่างกลุ่มของตัวแปรอิสระกับกลุ่มของตัวแปรตาม แต่เมื่อรวมทุกมิติ ก็อาจอธิบายความสัมพันธ์ระหว่างกลุ่มของตัวแปรอิสระกับกลุ่มของตัวแปรตามได้บางส่วนซึ่งอาจจะเป็นส่วนน้อย
4. Cancorr จะใช้ในการหาความสัมพันธ์ระหว่างกลุ่มของตัวแปรตามกับกลุ่มของตัวแปรอิสระพร้อมกัน ไม่เหมือนกับการวิเคราะห์การถดถอยพหุคูณ (multiple regression analysis) ที่ใช้หาความสัมพันธ์ระหว่างตัวแปรตามเพียงตัวเดียวกับกลุ่มของตัวแปรอิสระ ไม่ผิดนักถ้าจะกล่าวว่า multiple regression analysis เป็นกรณีพิเศษของ canonical correlation analysis

   ในศาสตร์ทางด้านจิตวิทยา การศึกษาพฤติกรรมของมนุษย์ส่วนใหญ่มักจะเป็นการศึกษาเกี่ยวกับตัวแปรที่แสดงเหตุและผลหลากหลายอย่าง cancorr เป็นเทคนิคทางสถิติที่เหมาะสมที่จะใช้ในสถานการณ์เช่นนี้
5. หากมีกลุ่มของตัวแปรตามกับกลุ่มของตัวแปรอิสระ ผลของการวิเคราะห์การถดถอยพหุคูณระหว่างตัวแปรตามทีละหนึ่งตัวกับกลุ่มของตัวแปรอิสระหลายๆ ครั้งตามจำนวนตัวแปรตามที่มีในกลุ่มจะไม่เท่ากับผลของการวิเคราะห์โดยใช้ cancorr เพียงครั้งเดียว เพราะจะเพิ่มโอกาสที่จะมี Type I error (โอกาสที่จะปฏิเสธสมมติฐาน H_O ทั้ง ๆที่แท้จริงแล้ว H_O เป็นสมมติฐานที่ถูกต้อง) ยิ่งทดสอบซ้ำครั้งเท่าใด effective significance level จะเพิ่มขึ้น ทำให้ความเชื่อมั่นทางสถิติลดลง
6. สเกลที่ใช้วัดกลุ่มของตัวแปรอิสระอาจแตกต่างจากสเกลที่ใช้วัดกลุ่มของตัวแปรตาม ตัวอย่างเช่น ตัวแปรตามอาจเป็น Likert's scale 1-5 ในขณะที่ตัวแปรอิสระอาจจะเป็น binary ที่มีค่าได้เพียงสองค่า
7. ผลที่ได้จาก cancorr อาจไม่ช่วยในการกำหนดหรือพยากรณ์ค่าของกลุ่มตัวแปรตามหากเราทราบค่าของกลุ่มตัวแปรอิสระ ทั้งนี้ต้องยอมรับวัตถุประสงค์หลักของ cancorr ซึ่งมีไว้เพื่อหาความสัมพันธ์ระหว่างกลุ่มของตัวแปร 2 กลุ่ม ไม่เหมือนกับผลที่ได้จากการวิเคราะห์การถดถอยพหุคูณที่สามารถใช้ผลจากการเปลี่ยนแปลงในตัวแปรอิสระหนึ่งหน่วยในการกำหนดว่าค่าของตัวแปรตามจะเปลี่ยนแปลงไปมากน้อยเท่าใด

8. ในการวิเคราะห์ cancorr มีความเป็นไปได้สูงที่ canonical variates จะมีความสัมพันธ์กันสูง แต่ไม่สามารถอธิบายสัดส่วน variance ของกลุ่มตัวแปรดั้งเดิมในระดับที่สูงพอ
9. ในปัจจุบัน cancorr เป็นเพียงเทคนิคที่พัฒนาขึ้นสำหรับใช้ในเชิงพรรณนาหรือใช้ในการกลั่นกรองมากกว่าจะใช้ในการทดสอบสมมติฐานทางสถิติ (Tabachnick & Fidell , 2008)
10. จุดอ่อนที่สำคัญที่สุดของ cancorr ก็คือการตีความผลที่ได้ ซึ่งไม่ใช่เรื่องง่าย (Tabachnick & Fidell , 2008)

สมมติฐานที่จำเป็นสำหรับการวิเคราะห์ cancorr

1. ลักษณะการกระจายของตัวแปร (distribution) : ตัวแปรทั้งหลายในประชากรต้องมีการกระจายแบบ multivariate normal (นั่นหมายความว่า ตัวแปรทั้งหมดตลอดจน linear combination ของมันมีการกระจายแบบ Normal) โดยเฉพาะอย่างยิ่ง กรณีที่นักสถิติจำเป็นต้องมีการทดสอบผลทางสถิติ
   หมายเหตุ :- ในกรณีที่การกระจายของประชากรมีความเพี้ยนจาก multivariate normal แต่หากกลุ่มตัวอย่างมีขนาดใหญ่ การวิเคราะห์ cancorr ก็จะยังคงมีความน่าเชื่อถือ
2. ขนาดของกลุ่มตัวอย่าง : ต้องมีมากพอ โดยในขั้นต่ำควรมีขนาดกลุ่มตัวอย่าง 20 เท่าของตัวแปรที่ศึกษาอยู่ และจะเป็นการดีมาก ถ้าขนาดของกลุ่มตัวอย่างเป็น 40-60 เท่าของจำนวนตัวแปร (Barcikowski & Stevens,1975)
3. ข้อมูลที่มีปราศจากค่าผิดปกติหรือสุดโต่ง (outliers) หากในข้อมูลมีค่าเหล่านี้ จะมีผลต่อ correlation coefficient ค่อนข้างมาก
4. ความสัมพันธ์ระหว่างตัวแปรในแต่ละกลุ่มและระหว่างกลุ่มต้องเป็นเส้นตรง (linear)
5. ความผันผวนใน error term ต้องมีค่าคงที่ (homoscedasticity-ดู statistics talks #9) โดย variance ของแต่ละตัวแปรในกลุ่มหรือระหว่างกลุ่มต้องมีค่าคงที่ทุกระดับค่าของตัวแปรอื่น ๆ
6. เมทริกซ์ที่เป็น correlation matrix ต้องไม่มีลักษณะที่เป็น Ill-conditioning matrix ซึ่งจะเกิดขึ้นเมื่อตัวแปรหนึ่งมีความสัมพันธ์โดยสมบูรณ์ (perfect relationship) กับตัวแปรอื่น เป็นผลทำให้เราไม่สามารถหา inverted correlation matrix ตลอดจนไม่สามารถดำเนินการวิเคราะห์ cancorr ต่อไปได้

ข้อจำกัด

1. Linear combination ของกลุ่มตัวแปรหรือ variate ที่ได้อาจจะไม่มีความหมายในทางทฤษฎี และอาจมีความยุ่งยากพอสมควรในความพยายามที่จะตีความผลที่ได้

2.หากความสัมพันธ์ระหว่างกลุ่มของตัวแปรเป็น nonlinear อาจจะไม่สามารถหาความสัมพันธ์ได้

3.ผลที่ได้ขึ้นอยู่เป็นอย่างมากกับข้อมูลที่รวมเข้ามาหรือตัดออกไป

4.Correlation ไม่ได้บ่งบอกเหตุและผลโดยอัตโนมัติ

การใช้โปรแกรมสำเร็จรูปในการหา canonical correlation

ไม่สามารถใช้คำสั่งจาก menu แต่จะต้อง run โปรแกรมโดยการสั่งการในรูป syntax ซึ่งเป็นภาษาเฉพาะของโปรแกรมสำเร็จรูปนี้ ทั้งนี้สามารถเลือก run โปรแกรมได้สองอย่าง คือใช้ manova command หรือ cancorr command

ใช้ MANOVA syntax command

```
manova  Y1 , Y2 , Y3 with X1, X2 , X3 , X4
/discrim all apha(1)
/print= sig( eigen dim).
```

คำอธิบาย : manova เป็นคำสั่งใน syntax
Y1 , Y2 , Y3 .....เป็นกลุ่มตัวแปรตาม (dependent variables)
X1 , X2 , X3 , X4.....เป็นกลุ่มตัวแปรอิสระ(independent variables)
/ เครื่องหมายคั่นระหว่างประโยค
discrim หมายถึง discriminant analysis subcommand
all alpha(1.0) หมายถึงสั่งให้แสดงผลโดยไม่คำนึงถึงค่านัยสำคัญ
print คำสั่งให้แสดง output
sig ย่อมาจาก significant
eigen หมายถึง eigenvalue
. เครื่องหมายจุลภาคมีไว้ตอนท้ายของคำสั่งเสมอ
หมายเหตุ :- ใน manova dependent variables หมายถึงตัวแปรตาม
Covariate variables หมายถึงตัวแปรอิสระ
และเนื่องจากเงื่อนไข fixed output format นักวิจัยต้องไม่สลับตำแหน่งโดยเอากลุ่มตัวแปรอิสระขึ้นก่อน มิเช่นนั้นจะเกิดความสับสนเอง

ใช้ Cancorr syntax command

```
Include  'd:\statisticstalks\Canonical correlation.sps'.
Cancorr set1=X1, X2, X3/
                set2=Y1,Y2,Y3,Y4/.
```

คำอธิบาย :

บรรทัดแรกเริ่มต้นระบุไดรฟ์และโฟลเดอร์ที่เก็บไฟล์ Canonical correlation.sps ซึ่งเป็น syntax file และเป็นส่วนหนึ่ง

ของโปรแกรมสำเร็จรูปที่มาพร้อมกันเป็นแพคเกจ

Cancorr เป็นคำสั่งในโปรแกรมสำเร็จรูป

/ เครื่องหมายกั้นระหว่างข้อความ

set1 ตามด้วยเครื่องหมายเท่ากับและชื่อของตัวแปรใน set1 ทั้งหมด

set2 ตามด้วยเครื่องหมายเท่ากับและชื่อของตัวแปรใน set2 ทั้งหมด

. เครื่องหมายจุลภาคมีตอนท้ายของ syntax

หมายเหตุ :- cancorr ไม่ใช้ตัวแปรอิสระหรือตัวแปรตาม แต่เป็นตัวแปรชุดที่ 1 และ 2

ความหมายของศัพท์เฉพาะที่ใช้และศัพท์ทางสถิติ :

Variable : ตัวแปรซึ่งแบ่งออกเป็นสองกลุ่ม อาจแยกแยะออกเป็นกลุ่มตัวแปรอิสระ (Independent variable/ predictor Variable) และกลุ่มตัวแปรตาม (dependent variable / criterion variable) หรืออาจแยกแยะออกเป็นตัวแปรในกลุ่มที่หนึ่งและตัวแปรในกลุ่มที่สอง โดยจำนวนตัวแปรในแต่ละกลุ่มไม่จำเป็นต้องเท่ากัน

Canonical correlation analysis : การวิเคราะห์ความสัมพันธ์ระหว่างกลุ่มของตัวแปรสองกลุ่ม

Canonical variate / canonical variable : ตัวแปรสังเคราะห์ (synthetic variable / latent variable) ที่เกิดจาก linear combination ของกลุ่มตัวแปรแต่ละกลุ่ม โดยตัวแปรสังเคราะห์ของกลุ่มตัวแปรตามเรียกว่า dependent canonical variate / dependent canonical variable ส่วนตัวแปรสังเคราะห์ของกลุ่มตัวแปรอิสระเรียกว่า covariate canonical variate / covariate canonical variable

Canonical variates / variants ไม่ใช่ factor ในความหมายเดียวกับ factor (s) ที่ได้จากการทำ factor analysis โดยใน factor analysis จะมีการคำนวณหา factors ที่ทำให้ variance ระหว่างกลุ่ม (between-group variance)สูงสุด และ variance ภายในกลุ่ม (in-group variance) ต่ำสุด ในขณะที่ canonical variates คู่แรกจะเป็นการสร้าง linear combination จากกลุ่มตัวแปรอิสระและสร้าง linear combination จากกลุ่มตัวแปรตามที่มุ่งให้ความสัมพันธ์ระหว่าง variates สูงสุด จากนั้น canonical variates คู่ที่สองจะถูกสร้างขึ้นจาก residuals ที่เหลือจากvariates คู่แรกด้วยการสร้างlinear combination ของทั้งตัวแปรอิสระและตัวแปรตามในอีกมิติ (dimension) ที่มีความเป็นอิสระ หรือ

orthogonal กับมิติแรกโดยมุ่งให้ความสัมพันธ์ระหว่าง variate คู่ที่สองมีความสัมพันธ์สูงสุด เป็นขั้นตอนดังนี้ไปจนครบ variate คู่สุดท้ายซึ่งมีอันดับเท่ากับค่าต่ำสุดระหว่างจำนวนตัวแปรอิสระและตัวแปรตาม

Canonical correlation : มีได้หลายความหมาย

ความหมายแรก ; หมายถึงการวิเคราะห์แบบ canonical (Canonical correlation analysis)

ความหมายที่สอง : หมายถึง canonical function / canonical root / characteristic root ซึ่งอาจมีหลาย function / root แต่ละ function หรือ root จะประกอบไปด้วย variate หนึ่งคู่ ได้แก่ dependent canonical variate และ Covariate canonical variate

ความหมายที่สาม : หมายถึง canonical correlation coefficient ซึ่งแสดงความสัมพันธ์เชิงเส้นตรงระหว่าง dependent canonical variate กับ covariate canonical variate โดย canonical correlation ที่ได้จาก function / root แรกจะอธิบายความสัมพันธ์ระหว่างกลุ่มตัวแปรสองกลุ่มได้เป็นส่วนใหญ่

Cancorr coefficient ที่ได้จาก variates จะมีขนาดเล็กลงๆ ดังนั้นจึงมักนิยมรายงานเฉพาะ correlation ที่มีขนาดใหญ่ที่สุด (ของ variates คู่แรก)

Cancorr coefficient จะมีค่าอยู่ระหว่าง 0 และ 1 และไม่มีค่าเป็นลบเนื่องจากสเกลที่สร้างขึ้นจาก weights ของ linear combination ได้ผ่านกระบวนการ standardizationมาแล้ว

Canonical roots : root แต่ละ root แสดงมิติหนึ่งของความสัมพันธ์ระหว่างคู่ของตัวแปรสังเคราะห์ที่มาจากตัวแปรสองกลุ่ม โดย root แรกจะมีขนาดใหญ่สุดและให้ข้อมูลเกี่ยวกับความสัมพันธ์ระหว่างคู่ของตัวแปรสังเคราะห์มากที่สุด root ถัดไปจะแสดงมิติอีกมิติหนึ่งของความสัมพันธ์ระหว่างคู่ของตัวแปรสังเคราะห์ที่ไม่อาจอธิบายได้ด้วยมิติแรก root อันดับถัดไปก็จะแสดงให้เห็นความสัมพันธ์ระหว่างคู่ของตัวแปรสังเคราะห์ที่ไม่อาจอธิบายได้ด้วยมิติแรกและมิติที่สอง จำนวน root ที่มีแสดงให้เห็นจำนวนมิติจะมีไม่เกินจำนวนตัวแปรที่มีสมาชิกสังกัดน้อยกว่าเพื่อนในระหว่างตัวแปรสองกลุ่ม นักวิจัยไม่จำเป็นต้องพิจารณาความสัมพันธ์ระหว่างคู่ของตัวแปรสังเคราะห์ในทุกมิติ โดยจะจำกัดตัวเองอยู่ในมิติที่สำคัญ ๆ เท่านั้น (หมายเหตุ :- มิติหนึ่งจะได้รับความสนใจต่อเมื่อ R_C =.30 หรือมากกว่า)

Canonical function : ได้แก่กลุ่มสัมประสิทธิ์ที่เกิดจาก linear combination ของตัวแปรที่มาจากกลุ่มตัวแปรอิสระและคัวแปรที่มาจากกลุ่มตัวแปรตามที่ผ่านกระบวนการ standardization อันเป็นผลทำให้ความสัมพันธ์ระหว่างตัวแปรมีมิติที่เป็นเอกลักษณ์ในตัวเอง แต่ละ function จะ orthogonal หรือมีความเป็นอิสระจาก function อื่นอันมีความหมายว่า Variate ของแต่ละ functionจะไม่มีความสัมพันธ์กับ variate ของ function อื่น

จำนวน function จะเท่ากับจำนวนสมาชิกของกลุ่มตัวแปรที่มีสมาชิกน้อยกว่า

Canonical weights / canonical function coefficients / canonical coefficients : แสดง partial correlation ระหว่างตัวแปรกับ canonical root ของมัน ช่วยให้สามารถทราบว่า แต่ละตัวแปรมีส่วนสำคัญมากน้อยเพียงใดในการกำหนดตัวแปรสังเคราะห์ และช่วยให้สามารถคำนวณค่าที่แท้จริงของ canonical variate ที่เรียกว่า canonical scores ได้

ความสำคัญอีกอย่างหนึ่งของ canonical weights ก็คือทำให้ผู้วิเคราะห์สามารถทราบองค์ประกอบหรือ “make-up” ของ root แต่ละ root ซึ่งอาจมีส่วนช่วยในการตีความผลที่ได้

Raw canonical coefficients for dependent (covariate) variables : บอกให้ทราบว่า หากตัวแปรในกลุ่ม (ไม่ว่าจะเป็นตัวแปรตาม / ตัวแปรอิสระ) เปลี่ยนไปหนึ่งหน่วยโดยตัวแปรอื่น ๆในกลุ่มมีค่าคงที่ จะมีผลทำให้ตัวแปรสังเคราะห์ / covariate ที่เกี่ยวข้องกับตัวแปรในกลุ่ม (ตัวแปรอิสระ / ตัวแปรตาม) เปลี่ยนไปมากน้อยเท่าใด

Standardized canonical function coefficients for dependent (covariate) variables : คือน้ำหนัก (weights) ของตัวแปรดั้งเดิมจากกลุ่มตัวแปรสองกลุ่ม (ตัวแปรอิสระหรือตัวแปรตาม/กลุ่มตัวแปรที่หนึ่งหรือกลุ่มตัวแปรที่สอง) ใน linear combination ที่ใช้ในการสร้าง canonical variateที่มีความสัมพันธ์กัน น้ำหนักนี้จะถูกกำหนดให้ canonical correlation ระหว่างสอง variates มีค่าสูงสุด

Canonical factor loadings/ canonical loadings / structure correlation coefficients / factor structure / structure coefficients : บอกความสัมพันธ์ระหว่างตัวแปรสังเคราะห์กับตัวแปรแต่ละตัวในกลุ่มของมัน (กรณี regular loadings หรือ กรณีปกติ) หรือบอกความสัมพันธ์ระหว่างตัวแปรสังเคราะห์กับตัวแปรแต่ละตัวที่อยู่ข้ามกลุ่ม (กรณี cross loadings)

Canonical loadings นอกจากจะให้ข้อมูลความสำคัญของตัวแปรนั้น ๆใน canonical solution แล้ว ยังทำให้เห็นความเชื่อมโยงระหว่างกลุ่มของตัวแปรตามและกลุ่มของตัวแปรอิสระผ่าน canonical root

จาก loadings ที่ได้ เราอาจต้องมีการกำหนดชื่อของตัวแปรสังเคราะห์ เพื่อแสดงมิติของความสัมพันธ์และความหมายที่มี แต่ในบางครั้งอาจทำไม่ได้หรือค่อนข้างเป็นเรื่องที่ยาก

Canonical scores : ค่าของ canonical variable ของแต่ละ case หาได้จากการเอา canonical coefficients (ของแต่ละตัวแปร) คูณกับ standardized scores ของตัวแปรแต่ละตัวแปรใน case หนึ่งๆและรวมผลเข้าด้วยกันเป็น canonical scores สำหรับ case นั้น ๆ

Canonical correlation coefficient (R_c) : Pearson's correlation coefficient ระหว่างตัวแปรสังเคราะห์สองตัว (dependent canonical variate และ covariate canonical variate) ที่ได้จาก canonical function หนึ่งๆ หากยกกำลังสองจะบอกสัดส่วนของความผันผวนใน variate หนึ่งที่สามารถอธิบายได้ด้วยอีก variate หนึ่ง ใช้บอกระดับตลอดจนทิศทางของความสัมพันธ์ระหว่าง variate แต่ละคู่

Squared canonical correlation : สัดส่วน variance ที่มีร่วมกันระหว่างตัวแปรสังเคราะห์ทั้งสองใน function หนึ่งๆ และบอกสัดส่วนของ variance ที่มีร่วมกันระหว่างตัวแปรสองกลุ่ม (อิสระ / ตาม) หรืออีกนัยหนึ่งใช้บอกสัดส่วน variance ของ one group variate ที่อธิบายได้ด้วย group variate อีกอัน

Redundancy : สัดส่วน variance ของตัวแปรดั้งเดิมในกลุ่มหนึ่งๆที่อธิบายได้ด้วย variate จากอีกกลุ่มหนึ่ง

Redundancy coefficients หรือ redundancy index : ใช้วัดสัดส่วนเป็นร้อยละของ variance ของตัวแปรดั้งเดิมจากกลุ่มที่หนึ่ง (กลุ่มตัวแปรอิสระ / กลุ่มตัวแปรตาม) ที่สามารถอธิบายได้โดยตัวแปรดั้งเดิมของอีกกลุ่มหนึ่ง (กลุ่มตัวแปรตาม/ กลุ่มตัวแปรอิสระ) โดยจะมี redundancy coefficient สองจำนวนต่อ canonical correlation หนึ่งๆ (redundancy coefficient ของ covariate canonical variate ที่ใช้พยากรณ์ variance ของกลุ่มตัวแปรตาม และ redundancy coefficient ของ dependent canonical variate ที่ใช้พยากรณ์ variance ของกลุ่มตัวแปรอิสระ)

Pooled redundancy coefficients : ผลรวมของ redundancy coefficients ของตัวแปรทุกตัวในกลุ่ม (ไม่ว่าจะเป็นกลุ่มตัวแปรอิสระหรือกลุ่มตัวแปรตาม) จะบอกประสิทธิภาพของ canonical variate ทุกตัวในการ capture variance ของตัวแปรเดิม (original variables)

Canonical communality coefficients : ผลรวมของ sq. structure coefficients ของ canonical variables ทั้งหมดของตัวแปรหนึ่งๆ จะบอกประโยชน์หรือความสำคัญของแต่ละตัวแปรในการวิเคราะห์ cancorr หากตัวแปรใดมี

canonical community ต่ำ หมายความว่าแบบจำลองที่ใช้อยู่ล้มเหลวและผู้วิเคราะห์อาจพิจารณาตัดสินใจเอาตัวแปรตัวนั้นออกจากกระบวนการวิเคราะห์ cancorr

Canonical variate adequacy coefficients : ค่าเฉลี่ยของ sq. structure coefficient ของกลุ่มตัวแปรหนึ่งๆในแต่ละ function จะบอกว่า canonical function แต่ละฟังก์ชั่นสามารถอธิบาย variance ของ dependent variable หรือ independent variable ได้มากน้อยเท่าใด

Pillai's trace (Pillais) : ใช้เพื่อทดสอบ null hypothesis (H_O : Canonical correlations =0) หรือ อีกนัยหนึ่งก็คือ ไม่มีความสัมพันธ์เชิงเส้นตรงระหว่าง variates ที่มาจากตัวแปรสองกลุ่ม) คำนวณหาได้โดยหาผลรวมของ squared canonical correlations

Hotelling-Lawley trace (Hotellings) : ใช้เพื่อทดสอบ null hypothesis เฉกเช่น Pilllais สามารถคำนวณได้โดยหาผลรวมของ (canonical correlation^2 ÷ (1- canonical correlation^2))

Wilks' lambda (Wilks) : ใช้เพื่อทดสอบ null hypothesis เฉกเช่น Pillais และ Hotelling คำนวณได้จากผลคูณของ (1- canonical correlation^2) ทุกตัว

Roy's greatest root (Roys) : ใช้เพื่อทดสอบ null hypothesis เฉกเช่น Pillais Hotelling และ Wilks คำนวณได้จากสูตร (eigen value ที่มีค่าสูงสุด ÷ (1+eigenvalueที่มีค่าสูงสุด)) เนื่องจากใช้ค่าสูงสุด ดังนั้นหากผลที่ได้จากการทดสอบสมมติฐานด้วยค่าสถิติสามตัว (Pillais / Hotellings / Wilks) บ่งบอกว่าไม่มีนัยสำคัญทางสถิติในขณะที่พบว่า Roys บ่งบอกว่ามีนัยสำคัญทางสถิติ เราจะสรุปว่า ผลการทดสอบไม่มีนัยสำคัญทางสถิติ

Approx.F : คือค่าสถิติ F ที่ได้โดยประมาณการ สำหรับการทดสอบที่เป็น multivariate test

Hypoth.DF/Error DF : degree of freedom คำนวณจาก mean squared errors ใช้ในการกำหนดค่าของ F ในบางครั้ง degree of freedom อาจไม่อยู่ในรูปเลขจำนวนเต็ม (integer) เนื่องจาก mean squared errors บ่อยครั้งอาจไม่อยู่ในรูปเลขจำนวนเต็ม

Roots : Roots หรือ dimensions จะมีจำนวนไม่เกินจำนวนของกลุ่มตัวแปรที่มีสมาชิกน้อยกว่า แต่ละ root ก็จะมี correlation ที่แสดงความสัมพันธ์ของ variate แต่ละคู่ตลอดจน eigenvalue root แรกจะมีขนาดใหญ่ที่สุดเนื่องจากให้ข้อมูลได้มากกว่าเพื่อน ตามด้วย root ที่สองที่ให้ข้อมูลน้อยกว่าลดหลั่นกันลงไปตามลำดับ root

Root no. : อันดับของroot ไล่เรียงจากค่า eigenvalue ที่สูงลงไปหาค่าที่ต่ำ จำนวน root=min(จำนวนตัวแปรในชุดที่ 1, จำนวนตัวแปรในชุดที่ 2)

Eigenvalue : ขนาดของ eigenvalue จะสะท้อน variance ใน canonical variates ที่อธิบายได้ด้วย canonical correlation ของมัน สามารถคำนวณได้จาก squared correlation ÷ (1-squared correlation)

Pct. : ร้อยละของ variance ใน canonical variates ที่อธิบายได้ด้วย canonical correlation ของมัน

Wilks L : เป็นการทดสอบโดยใช้ Bartlett's Chi-square (ดูจากค่า Wilks) เป็นการทดสอบสมมติฐานทางสถิติที่แตกต่างจาก Wilk's lambda ใน multivariate หรือ omnibus test (ซึ่งใช้ F-test) โดยการทดสอบนี้ไม่ได้ใช้เพื่อทดสอบความมีนัยสำคัญของ canonical correlation (หรือทดสอบว่า eigenvalue มีค่าต่างจากศูนย์) ทีละค่า แต่ใช้ในการทดสอบว่า ค่า canonical correlation ทั้งหมดยกเว้น canonical correlation ก่อนหน้าที่มีค่าใหญ่กว่าแตกต่างจากศูนย์อย่างมีนัยสำคัญหรือไม่ คำนวณได้จากผลคูณของ (1-canonical correlation^2) การทดสอบจะมีลักษณะเป็นลำดับขั้นตอน (sequential) โดยสมมติว่า เรามี canonical root อยู่ 3 ตัว (ขนาดของกลุ่มตัวแปรที่มีจำนวนสมาชิกน้อยที่สุด = 3) และถ้ากำหนดว่า canonical correlation ของ root ที่ 1-3 เป็น R_{C1} , R_{C2} , R_{C3} ตามลำดับ

จะมีการคำนวณดังนี้

Wilk L ของ root ที่ 1 = $(1-R^2_{C!})$. (1- R^2_{C2}) $(1-R^2_{C3})$

Wilk L ของ root ที่ 2 = $(1-R^2_{C2})$. (1- R^2_{C3})

Wilk L ของ root ที่ 3 = $(1-R^2_{C3})$

ก่อนจบเอกสารวิชาการนี้ มีประเด็นที่ต้องหยิบยกดังนี้ :-

ประเด็นที่ 1

ต้องเข้าใจว่า canonical correlation ไม่ได้ใช้วัดสัดส่วน varianceของตัวแปรดั้งเดิม (original variables) ที่สามารถอธิบายได้ด้วย canonical variate แต่ใช้บอกความสัมพันธ์ระหว่างผลรวมถ่วงน้ำหนักของตัวแปรสองกลุ่ม

ประเด็นที่ 2

เนื่องจาก canonical coefficients อาจได้รับอิทธิพลจากการมี multicollinearity ระหว่างตัวแปรในกลุ่ม มีผลทให้เครื่องหมายอาจแตกต่างไปจาก correlation กับ canonical variable ได้ ดังนั้นความสัมพันธ์ระหว่างตัวแปรกับ canonical variable ควรใช้ structure coefficient จะมีความเหมาะสมกว่า

ประเด็นที่ 3

มีความเป็นไปได้ที่ตัวแปรบางตัวมีค่า canonical weights ใกล้ 0 ในขณะที่ canonical factor loadings (structure coefficients) สูง โดยเฉพาะกรณีที่ตัวแปรใดตัวแปรหนึ่งมี variance ร่วมกันกับตัวแปรอื่น ๆ ทำให้ตัวแปรนั้นซ้ำซ้อน (redundant) กับตัวแปรอื่น ๆ มีผลทำให้ canonical weights เข้าใกล้ศูนย์ แต่ structure correlations อาจสูง

ประเด็นที่ 4

หากเปรียบเทียบความแตกต่างระหว่าง cancorr analysis กับ factor analysis (Statistics Talks #12-15) สามารถแจกแจง
ได้ดังนี้

Cancorr Analysis	Factor Analysis
1.มุ่งเน้นการแยกแยะตัวแปรออกเป็นตัวแปรอิสระและตัวแปรตาม	1.มุ่งศึกษาโครงสร้างของตัวแปรทั้งหมด
2.สร้างตัวแปรสังเคราะห์ (latent variable / variate) และมุ่งเน้นความสัมพันธ์ที่มีระหว่าง variate นี้	2.สร้างตัวแปรสังเคราะห์ขึ้นเช่นเดียวกันแต่ไม่ได้เน้นความสัมพันธ์
3.ศึกษากลุ่มตัวแปรต้นและตัวแปรตาม	3.ศึกษาความเป็นอิสระต่อกันและกัน

Contribution this issue: ดร.ดนัย ปัตตพงศ์

อยากเรียนรู้การนำสถิติข้างต้นนี้ไปใช้ในการวิจัยระดับสารนิพนธ์ (independent study) วิทยานิพนธ์ (thesis) ดุษฎีนิพนธ์(dissertation) ปรึกษาได้ที่ dpattaphongse@gmail.com

* ผู้แต่ง MBA's Made Easy (160+ issues) เอกสารวิชาการด้านศาสตร์การบริหารธุรกิจที่ช่วยให้ธุรกิจสามารถยืนหยัดและอยู่รอดได้ในภาวะที่โลกเปลี่ยนแปลงอยู่ตลอดเวลา

* ผู้พัฒนา FINALYSIS… a dedicated software สำหรับให้บริการนักธุรกิจที่ต้องการวิเคราะห์ความเป็นไปได้ทางการเงินของโครงการพัฒนาอสังหาริมทรัพย์ (บ้านจัดสรร/จัดสรรที่ดินเพื่อการอุตสาหกรรม/อาคารชุด/อาคารสำนักงานให้เช่า) โรงแรม โรงพยาบาลเอกชน ห้างสรรพสินค้า โรงงานน้ำตาล โรงงานกระดาษ โรงไฟฟ้าชีวมวล ฯลฯ ได้เห็นตัวเลขก่อนโครงการเกิด หลีกเลี่ยงความผิดพลาดเป็นร้อยเป็นพันล้านหากเกิดการลงทุนจริง(กำหนด DEBUT 1 เมษายน 2569)

* ผู้แต่งหนังสือ"การวิเคราะห์ความเป็นไปได้ทางการเงินและการจัดวงเงินเครดิตของโครงการลงทุน"ประกอบด้วยตัวอย่างของธุรกิจจริงที่ไม่เปิดเผยชื่อนับ 100 บริษัท ครอบคลุมอุตสาหกรรม 24 อุตสาหกรรม

*Co-developer ซอฟแวร์ en◉gex◉cel® สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)