20

Discriminant Analysis (DA)

เป็นเทคนิคทางสถิติที่ใช้ในการแยกแยะจัดกลุ่ม(classification) โดยใช้ตัวแปรอิสระ(independent variables/ discriminating variables/ predictor variables) ในการพยากรณ์ค่าของตัวแปรตาม (dependent variable / criterion variable / grouping variable)ซึ่งมีได้หลายค่า แต่ละค่าจะแสดงกลุ่มที่สังกัด หากค่าของตัวแปรตามมีได้เพียงสองค่า เราเรียกวิธีการในการแยกแยะจัดกลุ่มทางสถิตินี้ว่า discriminant analysis (DA) แต่ถ้าตัวแปรตามมีค่ามากกว่าสองค่า เราเรียกวิธีการในการแยกแยะจัดกลุ่มทางสถิตินี้ว่า multiple discriminant analysis (MDA)

เพื่อให้เห็นขอบเขตของการนำ DA ไปใช้ประโยชน์ในทางปฏิบัติ เราสามารถพิจารณาตัวอย่างต่อไปนี้

วงการทรัพยากรมนุษย์และการจ้างงาน

สายการบินนานาชาติขนาดใหญ่แห่งหนึ่งทำการเก็บรวบรวมข้อมูลเกี่ยวกับพนักงานสามประเภทได้แก่ บุคลากรที่ให้การบริการลูกค้า (ground hostess/air hostess) ช่างเครื่อง และ พนักงานวางแผนเส้นทางการบิน เติมน้ำมัน และจัดให้มีการทดสอบทางจิตวิทยาที่ใช้เป็นมาตรวัดความสนใจในกิจกรรมกลางแจ้ง การเข้าสังคม และความเป็นคนอนุรักษ์นิยม

วงการนักชีววิทยา

นักชีววิทยาใช้ DA ช่วยแยกแยะประเภทของดอก iris สามชนิดโดยพิจารณาจากตัวแปรสี่ตัวได้แก่ความกว้างของกลีบดอกไม้ ความยาวของกลีบดอกไม้ ความกว้างของกลีบเลี้ยง และความยาวของกลีบเลี้ยง

วงการตรวจสอบธนบัตรปลอม

เจ้าหน้าที่สามารถใช้ DA ผ่านเครื่อง scan เพื่อตรวจสอบธนบัตรของจริงแยกจากธนบัตรปลอมโดยพิจารณาจากความยาว ความกว้างด้านขวา ความกว้างด้านซ้าย ขอบบน ขอบล่าง เส้นทแยงมุมของพื้นที่ ๆมีการพิมพ์

วงการกีฏวิทยา

นักศึกษาด้านกีฏวิทยาสามารถใช้ DA ในการกำหนดชนิดของแมลงประเภทด้วงที่มีสองชนิดโดยอาศัยตัวแปรที่ประกอบไปด้วย ความกว้างของขาคู่หน้า ความกว้างของขาคู่หลัง และความกว้างของอวัยวะสืบพันธ์

วงการสถาบันการเงิน

ใช้ (Linear) DA ในการประเมินความเสี่ยงของลูกค้าผู้กู้ที่มี profile ไม่แตกต่างมากจากกลุ่มลูกค้าทั่วไป และใช้ (quadratic) DA แทนสำหรับกลุ่มลูกค้าที่ profile มีลักษณะเฉพาะ

วงการการแพทย์

แพทย์สามารถใช้ DA ในการกำหนดแบ่งคนไข้ที่มีความเสี่ยงสูงที่จะประสบภาวะหัวใจล้มเหลวแยกออกจากคนไข้ที่มีความเสี่ยงต่ำ โดยพิจารณาจากลักษณะข้อมูลส่วนบุคคล (ระดับคลอเรสเตอรอล ดัชนีมวลกาย) พฤติกรรมการออกกำลังกาย (จำนวนกี่นาทีต่อสัปดาห์) พฤติกรรมการสูบบุหรี่ (จำนวนซองต่อสัปดาห์)

วงการการแพทย์

ในการวินิจฉัยทางการแพทย์ (Linear) DA จะใช้ในการวิเคราะห์ชนิดของฝีที่อาจก่อมะเร็งในหมู่คนไข้ที่โครงสร้างของฝีในกลุ่มคนไข้มีความคล้ายคลึงกันและใช้ (Quadratic) DA ในการวิเคราะห์ชนิดของฝีที่มีโครงสร้างแตกต่างจากที่พบในกลุ่มคนทั่วไป

วงการสินเชื่อสถาบันการเงิน

สถาบันการเงินใช้ DA ในการตัดสินว่า ควรจะมีการอนุมัติสินเชื่อเงินกู้ซื้อรถยนต์หรือไม่โดยอาศัยข้อมูลจากอายุ รายได้ สถานะสมรส หนี้ที่มีอยู่ การมีบ้านที่อยู่ของตนเองของกลุ่มลูกค้าที่จ่ายชำระหนี้ตรงตามกำหนดในอดีต และกลุ่มของลูกค้าที่บิดพลิ้วชำระหนี้

วงการการศึกษา

สถาบันการศึกษาใช้ DA ในการพยากรณ์นักศึกษาที่จะสอบผ่านหรือล้มเหลวในการเรียนวิชาหนึ่งๆ โดยพิจารณาจากจำนวนเวลาที่ใช้ศึกษาทบทวนต่อสัปดาห์ ระดับความวิตกกังวลเกี่ยวกับการสอบ และอัตราการเข้าชั้นเรียน

วงการตรวจสอบที่มาของผลิตภัณฑ์

นักโบราณคดีใช้ DA ในการระบุว่า เศษของเครื่องปั้นดินเผาที่ขุดพบมาจากแหล่งโบราณคดีแหล่งใด โดยอาศัยข้อมูลจากแล็ปในการตรวจสอบองค์ประกอบของอลูมิเนียม เหล็ก แมกนีเซียม แคลเซียม โซเดี่ยม

สมมติฐานที่จำเป็นเพื่อให้การวิเคราะห์ทางสถิติมีความน่าเชื่อถือ

1)ความสัมพันธ์ระหว่างตัวแปรอิสระด้วยกันต้องเป็นเส้นตรง

2)ในแต่ละกลุ่มตัวแปรอิสระต้องมีลักษณะเป็นmultivariate normality

3)ตัวแปรอิสระต้องมี covariance matrices เท่ากันทุก ๆ กลุ่ม(homogeneity of variance-covariance matrices)

4)ขนาดของกลุ่มตัวอย่างของกลุ่มที่เล็กกว่าต้องมีจำนวนมากกว่าจำนวนของตัวแปรอิสระ

5)ต้องไม่มีปัญหา multicollinearity ในระหว่างตัวแปรอิสระ

DA กับ Multivariate Analysis of Variance (MANOVA)

การวิเคราะห์สถิติโดยใช้ DA จะคล้ายคลึงกับ MANOVA ยกเว้นว่าตัวแปรตาม และ ตัวแปรอิสระจะมีการสลับที่กัน โดยใน DA เราจะมีการสร้างสมการที่ใช้ในการพยากรณ์เพื่อแยกแยะสมาชิกในกลุ่มหนึ่งออกจากอีกกลุ่มหนึ่ง แต่ใน MANOVA เราต้องการดูว่า สมาชิกในแต่ละกลุ่มจะแตกต่างจากกันในมาตรวัดที่มีหลายๆอย่างมากน้อยแค่ไหน เนื่องจากมีความคล้ายคลึงกันระหว่าง DA และ MANOVA สมมุติฐานที่จำเป็นต้องมีเพื่อให้การวิเคราะห์ทางสถิติที่ใช้ DA หรือ MANOVA จึงเหมือนกัน

DA กับ Logistic regression

ทั้ง DA และ logistic regressionเป็นเทคนิคทางสถิติที่ใช้ในการแยกแยะจัดกลุ่ม โดยตัวแปรอิสระอาจเป็นได้ทั้งตัวแปรแยกประเภท (categorical variable) หรือตัวแปรที่มีค่าต่อเนื่อง (continuous variable)

เมื่อใดก็ตามที่มีข้อสงสัยว่าข้อสมมติฐานเกี่ยวกับ multivariate normality ไม่น่าจะถูกต้อง เราต้องใช้ logistic regression อนึ่งหากตัวแปรทุกตัวเป็นตัวแปรที่มีค่าต่อเนื่อง เราควรใช้ DA แทน logistic regression และเมื่อใดก็ตามที่ตัวแปรอิสระมีค่าได้เพียงสองค่า เราไม่ควรใช้ DA ยกเว้นว่าตัวแปรตามสามารถแบ่งออกเป็นสองกรณีที่มีโอกาสเกิด 50-50เท่า ๆกัน และหากขนาดของกลุ่ม มีความแตกต่างกันค่อนข้างมาก การใช้ logistic regression จะเหมาะกว่าการใช้ DA แต่ถ้าเมื่อใดก็ตามที่ สมมติฐานของ DA มีครบ จะมีpowerในการวิเคราะห์มากกว่า logistic regression เนื่องจากโอกาสที่จะมี Type II errors กล่าวคือ ไปให้การยอมรับ null hypothesis ที่ผิดจะมีโอกาสต่ำกว่า

DA กับ การวิเคราะห์การถดถอยทวีคูณ (Multiple regression analysis)

ใน multiple regression เราใช้ตัวแปรอิสระในการพยากรณ์ค่าของตัวแปรตามที่มีค่าต่อเนื่อง ใน DA เราใช้ตัวแปรอิสระที่มีค่าต่อเนื่องหรือเป็นตัวแปรเชิงคุณภาพ (qualitative) ในการพยากรณ์ค่าของตัวแปรตาม โดยที่ตัวแปรตามมีลักษณะเป็นตัวแปรเชิงคุณภาพ ในDA เองก็มี algorithm ที่ให้สามารถเลือกตัวแปรอิสระที่จะนำเข้ามาเพื่อใช้ในการพยากรณ์คล้ายคลึงกับการทำ stepwise regression

วิธีการของ DA

จากข้อมูลที่มีการจัดกลุ่มไว้ให้แล้ว DA จะสร้างกฎเกณฑ์ที่ใช้ในการแยกแยะกลุ่มออกโดยใช้เทคนิคที่เรียกว่า Fisher's linear discriminant function analysis ทั้งนี้ถ้าสมมติว่าเรามีมาตรวัดที่แสดงด้วยตัวแปรอิสระ q ตัว (X_1 , X_2 ,…….. X_q) DA จะหา function ที่เป็น linear transformation ของ X1 , X2 ,…….. Xq ที่อยู่ในรูป

$$Z = a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots\ldots + a_q X_q$$

โดยที่ $\overline{Z}_1$ และ $\overline{Z}_2$ เป็นค่า group means ที่ transform แล้ว มีระยะห่างระหว่างกลุ่ม(between-groups) สูงสุดเมื่อเปรียบเทียบกับความผันผวนภายในกลุ่ม (within-group)ที่อยู่ในรูป Z-scale หากสมาชิกใดมี discriminant score ใกล้ $\overline{Z}_1$ มากกว่า $\overline{Z}_2$ จะถูกจัดเข้ากลุ่มหนึ่ง หากไม่ใกล้กว่าจะถูกจัดเข้ากลุ่มสอง ทั้งนี้ classification rule ที่ใช้คือค่า cut-off point แสดงได้ด้วย

$$Z^* = (\overline{Z}_1 + \overline{Z}_2)/2$$ และถ้า $Z > Z^*$ จะถูกจัดอยู่อยู่กลุ่มหนึ่ง หากต่ำกว่าจะถูกจัดอยู่ในอยู่กลุ่มสอง

ขั้นตอนการวิเคราะห์ของ DA

จะแบ่งออกเป็น 2 ขั้นตอน

(1) ใช้ F-test (Wilks' lambda) เพื่อดูว่า discriminant model โดยรวมมีนัยสำคัญหรือไม่

(2) ในกรณีที่มีนัยสำคัญ จะมีการประเมินตัวแปรอิสระแต่ละตัวเพื่อดูว่าค่าเฉลี่ยแต่ละกลุ่มแตกต่างกันอย่างมีนัยสำคัญหรือไม่ เพื่อใช้ในการแยกแยะกลุ่มของตัวแปรตาม

ในกรณีของ 2-group DA จะมี discriminant function(หรือที่เรียกว่า canonical root) เพียงอันเดียว ในกรณีของ MDA จำนวน discriminant function (หรือที่เรียกว่า dimension) จะเท่ากับค่าที่น้อยกว่าระหว่าง g-1 และ p โดยที่ g= จำนวนกลุ่มของตัวแปรตาม ส่วนp= จำนวนของตัวแปรอิสระ โดยที่ discriminant function แต่ละอันจะ orthogonal (ไม่มีความสัมพันธ์)กัน ทั้งนี้ discriminant function อันแรกจะ maximize ความแตกต่างระหว่างค่าเฉลี่ยของกลุ่มที่อยู่ในตัวแปรตาม function ที่ 2 จะ maximize ความแตกต่างระหว่างค่าเฉลี่ยของกลุ่มที่อยู่ในตัวแปรตาม โดย control factor ตัวแรกไว้

ในกรณีของ MDA เราต้องทำ pairwise group comparisons ระหว่างกลุ่ม (กรณี DA ซึ่งมี function เดียว ไม่สามารถกระทำได้) มีการทดสอบโดยใช้ F-test วัดระยะห่างที่เรียกว่า Mahalanobis distance ระหว่างค่าเฉลี่ยของแต่ละกลุ่ม (group means)

ความหมายเฉพาะที่ใช้ใน DA

Wilk's lambda: หรือบางทีเรียกว่า U statistic ใช้ในการทดสอบว่า มีตัวแปรอิสระใดที่มีความสำคัญต่อdiscriminant functionที่ได้ โดยถ้า

ค่า lambda ยิ่งเล็ก ยิ่งบ่งบอกความสำคัญของตัวแปรอิสระนั้น (ปกติ Lambdaจะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยถ้าเท่ากับ 1 หมายความว่าค่าเฉลี่ยของทุกกลุ่มเท่ากัน ส่วนค่าที่ต่ำกว่าแสดงว่าค่าเฉลี่ยของกลุ่มมีความแตกต่าง) และหากพบว่าตัวแปรอิสระเพียงตัวเดียวมีนัยสำคัญหมายความว่าโดยรวม model ที่ใช้มีนัยสำคัญ

Box's M test of equality of Covariance Matrices: ใช้ทดสอบเพื่อดูว่า covariance matricesของกลุ่มต่าง ๆมีความแตกต่างกันหรือไม่ อย่างไรก็ตามในกรณีที่ขนาดของกลุ่มตัวอย่างมีขนาดใหญ่ ความแตกต่างใน covariance matrices แม้จะเพียงเล็กน้อยอาจจะทำให้ Box's M test มีนัยสำคัญ

หากพบว่า Log Determinants ของกลุ่มทั้งสองมีค่าใกล้เคียงกัน เราจะมองข้ามผลการทดสอบโดย Box's M test และในส่วนของ menu-classification เราต้องเลือก separate-groups หรือไม่ก็ต้อง run model โดยใช้ within-groups และ separate groups covariance เปรียบเทียบดูว่าผลใกล้เคียงกันหรือไม่ ถ้าใกล้เคียงกันเราจะมองข้ามผลการทดสอบ Box's M test ที่บ่งบอกว่ามีนัยสำคัญทางสถิติ

Eigenvalues : บางทีเรียกว่า characteristic roots บ่งบอกสัดส่วนความสำคัญของ discriminant functions (ใช้ไม่ได้ หากเป็น 2 group DA เพราะมี discriminant function เดียว)

Canonical correlation : (ใช้สัญลักษณ์ R_C หรือ R*) ใช้วัดความเชื่อมโยงระหว่างกลุ่มที่แสดงด้วยตัวแปรตาม และ discriminant function โดยจะมีหนึ่ง canonical correlation ต่อหนึ่ง discriminant function หากมีค่าเป็นศูนย์ แสดงว่าไม่มีความสัมพันธ์ระหว่างกลุ่ม หากมีค่าสูงแสดงว่ามีความสัมพันธ์สูงในระหว่างกลุ่มและdiscriminant function

R^2_C บ่งบอกความผันผวนในตัวแปรตาม ที่แยกแยะจัดเข้ากลุ่มด้วยตัวแปรอิสระ และ R_C ยังแสดงความสัมพันธ์ระหว่าง discriminant function กับ discriminant scores (ถ้าค่าเข้าใกล้ 1 หมายความว่าความผันผวนใน discriminant scores เกือบทั้งหมดสามารถอธิบายได้โดยความแตกต่างในระหว่างกลุ่ม)

ในกรณี 2-group DA , canonical correlationก็คือค่าPearson correlation ที่แสดงความสัมพันธ์ระหว่าง discriminant scores กับตัวแปรตาม

Model Wilks' lambda : ใช้ในการทดสอบ discriminant function โดยรวมว่ามีนัยสำคัญหรือไม่หรืออีกนัยหนึ่งใช้เพื่อทดสอบว่า โดยรวม model มีนัยสำคัญหรือไม่(Null hypothesis: กลุ่มทั้งสองกลุ่มหรือมากกว่ามี mean discriminant function scores เท่ากัน)

Model Wilk's lambda ในที่นี้จะต้องไม่ใช้สับสนปะปนกับ Wilks' lambda ที่ใช้ทดสอบ test of equality of group means

Standardized Canonical Discrimination Function Coefficients : ทำหน้าที่คล้ายๆ beta weight ใน multiple regression โดยจะบอกความสำคัญของตัวแปรอิสระในการแยกแยะความแตกต่างระหว่างกลุ่มของตัวแปรตาม ทั้งนี้ความสำคัญนี้ขึ้นอยู่กับโมเดลที่ใช้ หากเราตัดตัวแปรอิสระบางตัว ค่าcoefficients นี้จะเปลี่ยนแปลงมาก ยิ่งถ้าเป็น MDA ค่า coefficient นี้จะไม่บอกว่ากลุ่มไหนที่ตัวแปรอิสระสามารถแยกแยะได้ต่ำหรือสูงมากกว่า ต้องใช้ group centroids หรือ factor structure แทนเท่านั้น

Structure coefficients /structure correlations/discriminant loadings/canonical structure matrix/factor structure matrix :แสดงความสัมพันธ์ระหว่าง discriminant scores กับ predictor variables โดยค่า coefficients ที่มีค่าสูงกว่า 0.3 ถือว่ามีความสำคัญ structure matrix coefficients ทำหน้าที่คล้ายคลึงกับ factor loadings ใน factor analysis(ดู Statistics talks # 12-15) ในกรณีของ MDA structure matrix coefficients จะช่วยผู้วิเคราะห์ให้สามารถตีความและกำหนดชื่อของ discriminant function แต่ละอันได้ง่ายขึ้น

Canonical Discriminant function coefficients :ใช้สร้าง prediction equation ซึ่งเป็น linear combination ของตัวแปรอิสระ

Functions at group centroids: แสดง mean discriminant scores สำหรับแต่ละกลุ่มของตัวแปรตาม สำหรับmodel ที่สามารถแยกแยะได้เด่นชัด group centroids (mean discriminant scores) จะต้องอยู่ห่างจากกัน หากอยู่ใกล้กัน โอกาสที่จะ misclassify มีสูง

Functions at group centroids ใช้ในการกำหนด cutting point ที่จะแยกแยะตัวอย่างแต่ละตัวอย่างเข้ากลุ่ม หากสองกลุ่มมีขนาดเท่ากัน cutting point จะเป็นค่ากึ่งกลางระหว่างค่าของ functions at group centroids หากขนาดของกลุ่มสองกลุ่มไม่เท่ากัน cutting point จะเป็นค่าเฉลี่ยถ่วงน้ำหนักของค่า functions at group centroids

Histogram: ใน DA ที่สามารถแยกแยะจัดกลุ่มได้ดี histogram จะแสดงกราฟแท่งทรงสูงล้อมรอบค่าเฉลี่ยโดยจะมีกราฟแท่งทรงต่ำอยู่ส่วนปลายของ histogram ทั้งสองด้าน

Classification results/classification matrix/confusion/assignment/prediction matrix : ใช้ในการประเมินความสามารถในการแยกแยะจัดกลุ่มของ DA หากการพยากรณ์โดย DA สมบูรณ์แบบ ทุก ๆกรณีจะอยู่บนแนวเส้นทแยงมุม สัดส่วนของกลุ่มตัวอย่างที่อยู่บนเส้นทแยงมุมจะแสดงร้อยละที่มีการแยกแยะจัดกลุ่มได้ถูกต้อง และเราเรียกตัวเลขร้อยละนี้ว่า hit ratio

Contribution this issue: ดร.ดนัย ปัตตพงศ์

งานวิจัยที่อ้างอิงเอกสารวิชาการฉบับนี้

จันทิมา พัฒนเดช และคณะ “ ลักษณะหลักทรัพย์ที่กองทุนรวมเลือกลงทุนในตลาดหลักทรัพย์แห่งประเทศไทย”.Journal of Community Development Research,9(2),2016.

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)