

Multinomial Logistic Regression(MLR)

ชื่อเรียกทางสถิติของการวิเคราะห์นี้: มีชื่อเรียกหลายชื่อ ได้แก่ multinomial logistic regression/ polytomous Linear regression / multiclass linear regression / softmax regression / multinomial logit / maximum entropy classifier/ conditional maximum entropy model

Multinomial logistic regression (MLR) เป็นเทคนิคที่ใช้ในการพยากรณ์ค่าของตัวแปรตามหนึ่งตัวจากค่าของตัวแปรอิสระมากกว่าหนึ่งตัวขึ้นไป โดยที่ตัวแปรตามจะเป็นตัวแปรที่เป็นนามบัญญัติ (nominal) ที่มีค่าได้ตั้งแต่สามค่าขึ้นไป โดยแต่ละค่าจะแทนประเภท/ชนิด/กลุ่มของตัวแปรตาม ตัวอย่างเช่น ในการศึกษาทางเลือกของนักเรียนชั้นมัธยมศึกษาตอนปลาย ตัวแปรตามอาจมีค่าเป็น 1 ถ้านักเรียนคนนั้นต้องการเรียนต่อในระดับวิทยาลัย/มหาวิทยาลัย มีค่าเป็น 2 ถ้านักเรียนคนนั้นต้องการเรียนต่อในระดับอาชีวะ มีค่าเป็น 3 ถ้านักเรียนคนนั้นต้องการไปประกอบอาชีพ ส่วนตัวแปรอิสระที่เกี่ยวข้องในการพยากรณ์อาจเป็น เพศ เกรดเฉลี่ยสะสม อาชีพของบิดามารดา รายได้ของบิดามารดา ฯ เป็นต้น

เพื่อให้เข้าใจได้ง่ายขึ้น ลองมาพิจารณาตัวอย่างการใช้ MLR ในแวดวง ต่าง ๆ

วงการแพทย์ : ใช้ MLR ในการวิเคราะห์โรคใช้ตับอักเสบสองชนิด คือไวรัสตับอักเสบบชนิด B (hepatitis B virus- HBV) และไวรัสตับอักเสบบชนิด C (hepatitis C virus -HCV) ตัวแปรอิสระได้แก่ อายุ เพศ ถิ่นที่พำนัก (ชนบท/เมือง) ระดับการศึกษา สถานภาพสมรส สถานภาพการจ้างงาน ความเสี่ยงจากการทำงานด้านเกษตร ความเสี่ยงพิเศษ (เป็นพ่อเลี้ยงเดี่ยว มาจากสถานดูแลเด็กกำพร้า มีอาการติดยา มีพฤติกรรมรุนแรง เหลือค้ำมนุษย์ ค่าจ้างขั้นต่ำ ไม่สมประกอบ) ส่วนตัวแปรตามได้แก่ เป็น HBV หรือ HCV หรือไม่มีอาการติดเชื้อใด ๆ

วงการขายโทรศัพท์มือถือสมาร์ทโฟน : บริษัทขายโทรศัพท์สมาร์ทโฟนต้องการศึกษาความปรารถนาของลูกค้าที่ต้องการซื้อโทรศัพท์เครื่องใหม่ โดยใช้ MLR ในการตอบข้อสงสัยเกี่ยวกับความต้องการของผู้บริโภค ตัวแปรอิสระได้แก่ อายุ เพศ และราคา ส่วนตัวแปรตามได้แก่ โทรศัพท์ยี่ห้อแอปเปิ้ล ซัมซุง และอื่น ๆ

วงการสวนสัตว์ : นักชีววิทยาอาจสนใจในประเภทอาหารที่ระแฆเลือก(ปลา สัตว์ไม่มีกระดูกสันหลัง สัตว์เลี้ยงคลาน นก และ อื่น ๆ) โดยพิจารณาจากความยาว อายุ ถิ่นที่อยู่ซึ่งอาจเป็นทะเลสาบ หนองน้ำ ตัวแปรทางสภาพแวดล้อมอื่น ๆ เช่น อุณหภูมิของน้ำ ความหนาแน่นของเหยื่อ การมีสัตว์คู่แข่งที่แย่งอาหารกัน

วงการเครื่องดื่ม : นักโภชนาการในประเทศอังกฤษมีความสนใจศึกษาแนวโน้มของผู้คนเกี่ยวกับการบริโภคเครื่องดื่ม ซึ่งเปลี่ยนไปตาม generation ของผู้คนที่เปลี่ยนไป และได้ใช้ MLR เป็นเครื่องมือศึกษาแนวโน้มการบริโภคเครื่องดื่ม ตัวแปรตามได้แก่ กาแฟ/ชา/เครื่องดื่มปราศจากแอลกอฮอล์ / เครื่องดื่มคุณภาพสูงเช่นเบียร์สด ค็อกเทล และไวน์ ส่วนตัวแปรอิสระได้แก่ generation อายุ ถิ่นที่อยู่อาศัย (อากาศหนาวเย็นทางภาคเหนือ)

วงการดูแลสุขภาพ : นักวิจัยใช้ MLR ในการแยกแยะประชากรสุขภาพสตรีออกเป็นผู้ที่อ้วนมีน้ำหนักตัวค่อนข้างมาก ผู้ที่มีน้ำหนักตัวน้อยค่อนข้างผอมบาง และผู้ที่มีน้ำหนักปกติ ตัวแปรอิสระได้แก่ อายุ รายได้ของครอบครัว การศึกษา (ไม่รู้หนังสือ / อ่านออกเขียนได้) อาชีพ (บริการ/รับจ้าง/เป็นเจ้าของผู้ประกอบการ) ขนาดของครัวเรือน(<10, >=10) จำนวนผู้เยาว์ที่อยู่ร่วมด้วย ถิ่นที่อยู่(ชนบท/ในเมือง)

วงการการศึกษา : นักการศึกษาใช้ MLR ในการศึกษา การตัดสินใจเลือกประกอบอาชีพของนักศึกษาในสายงานต่าง ๆ โดยอาศัยข้อมูลเกี่ยวกับประชากรศาสตร์ (ระดับการศึกษาของนักศึกษา ภูมิหลัง เชื้อชาติ และอาชีพของบิดา สถานะภาพทางสังคม เศรษฐกิจ ทักษะ ประสบการณ์) ส่วนตัวแปรตามได้แก่ สายอาชีพ สายช่าง สายการบริหาร

วงการผลิตไอศกรีม : ใช้ MLR ในการพยากรณ์ว่า ผู้บริโภคจะเลือกทานไอศกรีมใด โดยตัวแปรอิสระได้แก่ อายุ เพศ แนวภาพยนตร์ที่ชื่นชอบ ส่วนตัวแปรตามได้แก่ ไอศกรีมวานิลลา ไอศกรีมช็อกโกแลต และ ไอศกรีมสตอเบอรี่

วงการโรงภาพยนตร์ : ใช้ MLR ศึกษาการเลือกชมภาพยนตร์ ตัวแปรตามได้แก่ประเภทภาพยนตร์ (ตลก แอ็คชั่น ซีวิต) ส่วนตัวแปรตามได้แก่ อายุ เพศ สถานภาพของคู่ควงที่มาด้วยกัน รายได้ ระดับการใช้สื่อสังคมออนไลน์ ระดับการศึกษา สถานภาพสมรส อาชีพ สถานที่ไปชมภาพยนตร์

วงการประกันสุขภาพ : บริษัทประกันทำการศึกษาการเลือกทำประกันสุขภาพของลูกค้าที่ทำประกันภัย ตัวแปรตาม ได้แก่ แผนการทำประกัน (แผน A /B /C เบี้ยประกัน ค่าใช้จ่ายที่สามารถหักลดหย่อนได้ ค่าใช้จ่ายที่แบ่งจ่าย) ตัวแปรอิสระ ได้แก่ อายุ เพศ จำนวนวันขาดงานก่อนมาทำประกันภัย รายได้ สุขภาพ ของครอบครัว สุขภาพของผู้เอาประกัน ตำแหน่งหน้าที่การงาน

วงการบริหารทรัพยากรมนุษย์ : นักวิจัยใช้ MLR ในการพยากรณ์ว่า พนักงานขององค์กรภาคธุรกิจทำงานอยู่ในระดับใด โดยตัวแปรอิสระ ได้แก่ ประสบการณ์ (จำนวนปี) ระดับการศึกษา (มัธยม ปริญญาตรี ปริญญาโท ปริญญาเอก) ผลประกอบการขององค์กร (อัตราการเติบโตของรายได้ ผลกำไร) ผลงานที่ผ่านมา ทักษะ ความเป็นผู้นำ ประสบการณ์ในอุตสาหกรรม ส่วนตัวแปรตาม ได้แก่ พนักงานระดับล่าง พนักงานระดับกลาง พนักงานระดับบน

วงการสถาบันการเงิน : โดยการใช้ MLR ควบคู่กับ machine learning algorithms และ pattern recognition technique ในการตรวจจับการโกงในแวดวงการเงิน ตัวแปรอิสระ ได้แก่ ชุกรกรรม (จำนวนเงินที่เกี่ยวข้อง/ความถี่/ สถานที่ / เวลา / พ่อค้าผู้มีส่วนเกี่ยวข้องด้วย) พฤติกรรม (การเข้าระบบและตรวจสอบตัวตน / ประวัติการทำธุรกรรมและแนวโน้ม / อุปกรณ์และการท่องเว็บ) ส่วนตัวแปรตาม ได้แก่ การทุจริตด้านบัตรเครดิต การปลอมประวัติระบุตัวตน และ การหลอกลวงทางอินเทอร์เน็ต

สมมติฐานที่จำเป็นเพื่อให้ผลการวิเคราะห์ถูกต้องตามหลักสถิติ

- 1) ตัวแปรอิสระอาจเป็นตัวแปรที่มีค่าต่อเนื่องหรือตัวแปรนามบัญญัติ
- 2) ตัวแปรตามเป็นตัวแปรนามบัญญัติที่สามารถแบ่งออกได้หลายประเภท(ตั้งแต่ 3 ขึ้นไป) ต้องเป็นประเภทที่แยกแยะออกจากกันโดยเด็ดขาด(mutually exclusive) และต้องครอบคลุมทุก ๆประเภท (exhaustive)
- 3) ข้อมูลที่เก็บได้มาต้องมีความเป็นอิสระ ไม่ใช่เกิดจากการทดลองซ้ำหลายครั้งจากกลุ่มตัวอย่างเดิม
- 4) ข้อมูลไม่จำเป็นต้องมีการกระจายแบบ Normal ไม่จำเป็นต้องมีความสัมพันธ์ในเชิงเส้นตรง และไม่จำเป็นต้องมี variances เท่ากัน
- 5) ต้องไม่มี multicollinearity ในระหว่างตัวแปรอิสระ

- 6) ความสัมพันธ์ระหว่างตัวแปรอิสระกับ logit transformation ของตัวแปรตามต้องเป็นเส้นตรง
- 7) ต้องไม่มีตัวเลขที่มีค่าอยู่นอกกลุ่ม (outlier)
- 8) เพื่อให้ผลการวิเคราะห์ทางสถิติมีความน่าเชื่อถือ Hosmer & Lemeshow กำหนดให้ขนาดของกลุ่มตัวอย่างขั้นต่ำต่อตัวแปรอิสระเท่ากับสิบ

เปรียบเทียบ Multinomial logistic regression กับ Discriminant analysis

- MLR ไม่ได้มีข้อสมมติฐานใด ๆ เกี่ยวกับ normality, linearity หรือ homogeneity of variance ของตัวแปรอิสระ แต่เป็นสมมติฐานที่จำเป็นก่อนการวิเคราะห์โดยใช้ discriminant analysis ด้วยเหตุนี้เอง MLR จึงได้รับความนิยมมากกว่า discriminant analysis โดยเฉพาะอย่างยิ่งเมื่อนักสถิติอาจมีข้อสงสัยว่าข้อมูลที่มีอาจไม่เป็นไปตามข้อสมมติฐาน
- ในทางปฏิบัติหากข้อมูลมีการกระจายแบบ multivariate normal มีvariance และ covariance ที่ homoscedastic ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นตรง เมื่อนั้นนักวิจัยควรใช้ discriminant analysis เพราะเป็นเทคนิคทางสถิติที่มีประสิทธิภาพมากกว่า โดยจะมีความแม่นยำในการพยากรณ์แยกแยะตัวแปรตามได้ดีกว่า MLR

วิธีการคำนวณหา solution ของ MLR

◎ MLR เป็นแบบจำลองที่เรียกว่า multi-equation model โดยหากกำหนดให้ตัวแปรตามแบ่งแยกออกได้ k ประเภท ในกระบวนการของ MLR จะสร้างสมการ logit ขึ้น (k-1) สมการ แต่ละสมการจะมีรูปแบบดังต่อไปนี้

$$\log \{ \text{prob}(y=1)/1-\text{prob}(y=1) \} = \beta_{01} + \beta_{11} X_1 + \beta_{12} X_2 + \beta_{13} X_3 + \dots + \beta_{1m} X_m \quad \text{สมการ (1)}$$

$$\log \{ \text{prob}(y=2)/1-\text{prob}(y=2) \} = \beta_{02} + \beta_{21} X_1 + \beta_{22} X_2 + \beta_{23} X_3 + \dots + \beta_{2m} X_m \quad \text{สมการ (2)}$$

.

โดยที่ y เป็นตัวแปรตามที่มีค่าแต่ละค่าจะบ่งบอกประเภท ($y = 1, 2, \dots, k$)

prob = probability

$\text{prob}(y=1) / 1 - \text{prob}(y=1)$ แสดง odd ratio ที่ค่าของ y จะเท่ากับ 1

β 's แสดงค่าสัมประสิทธิ์หน้าตัวแปรอิสระแต่ละตัว

X_{ij} = ค่าของตัวแปรที่ j ในสมการที่ i

สมการที่ (1) จะใช้ในการพยากรณ์ความน่าจะเป็นที่ตัวอย่างจะถูกจัดอยู่ในตัวแปรตามประเภทที่หนึ่ง

สมการที่ (2) จะใช้ในการพยากรณ์ความน่าจะเป็นที่ตัวอย่างจะถูกจัดอยู่ในตัวแปรตามประเภทที่สอง

จำนวนสมการทั้งหมดจะมี $(k-1)$ สมการ โดยสมการสุดท้ายจะใช้ในการพยากรณ์ความน่าจะเป็นที่ตัวอย่างจะถูกจัดอยู่ในตัวแปรตามประเภทที่ $(k-1)$

โดย default ประเภทของตัวแปรตามที่มีค่าตัวเลข (numeric score) สูงสุดจะถูกใช้เป็น reference group ที่ตัวอย่างจะถูกจัดเข้าอยู่ในกลุ่มสุดท้าย และค่าสัมประสิทธิ์หน้าตัวแปรของ reference group จะเป็นศูนย์ นี่คือเหตุผลที่ว่าทำไมในการหา solution จึงมีสมการเพียง $(k-1)$ สมการ ขาดสมการที่ใช้ในการจัดกลุ่มเข้า reference group ไปหนึ่งสมการซึ่งไม่จำเป็นต้องเขียนเพราะหน้าตัวแปรอิสระทุกตัวมีค่าสัมประสิทธิ์เป็นศูนย์หมด

นักสถิติสามารถเลือกตัวแปรตามกลุ่มที่หนึ่งเป็น reference group หรือตัวแปรตามกลุ่มอื่นในกลุ่มใดกลุ่มหนึ่งเป็น reference group ได้โดยตรงจาก menu

- MLR ค้นหา solution โดยใช้วิธีการที่เรียกว่า Maximum likelihood method (MLE) ซึ่งเป็นกระบวนการคำนวณกลับไปกลับมาเพื่อหาคำตอบ แต่ในบางครั้ง MLE อาจล้มเหลวกลางคันและไม่สามารถหาคำตอบต่อไปได้

- ในบางครั้ง MLE อาจให้ผลที่ไม่น่าเป็นไปได้ เช่นการเปลี่ยนแปลงในตัวแปรอิสระหนึ่งหน่วยอาจเป็นผลให้ odds ของการเกิดเหตุการณ์หนึ่งเปลี่ยนไปเป็น หลายพันหรือหลายล้าน นี่อาจมีสาเหตุมาจากปัญหาของการมี multicollinearity ในข้อมูล หรืออาจมาจากไม่มีข้อมูลของกลุ่มตัวอย่างอยู่ในแต่ละเซลล์ของตัวแปรอิสระ (จำนวนตัวอย่างในเซลล์ = 0) หรืออาจเกิดจากกรณีการมี complete separation ที่ตัวแปรอิสระหนึ่งหรือมากกว่ามีค่าอยู่ที่ระดับหนึ่ง

เมื่อใดก็ตามที่ค่า standard errors ของตัวแปรอิสระบางตัวมีค่ามากกว่า 2.0 เมื่อนั้นสื่อเป็นนัยให้เห็นปัญหาว่าการหา solution โดย MLE มีปัญหา

โปรดสังเกตว่า ตัวแปรอิสระที่ไม่มีค่าต่อเนื่องจะถือเป็น ปัจจัย (factor) ส่วนตัวแปรอิสระที่มีค่าต่อเนื่องจัดเป็น covariates

- วิธีในการเลือกตัวแปรอิสระเป็นแบบใส่พร้อมกันหรือโดยตรง (simultaneous หรือ direct entry)

การใช้โปรแกรมสำเร็จรูปใน MLR

สามารถทำได้โดยใช้ menu ที่เรียกว่า drop-down menu หรือโดยใช้ syntax command

Syntax command ที่ใช้ในการวิเคราะห์ MLR อยู่ในรูปต่อไปนี้

```
NOMREG MultinomialDV(BASE=LAST ORDER=ASCENDING) BY Factor WITH Covariate
/MODEL=Factor Covariate
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

คำอธิบาย

- NOMREG ย่อมาจาก nominal regression คือการ run regression ที่ตัวแปรตามเป็นนามบัญญัติที่แบ่งออกได้ตั้งแต่ 3 กลุ่ม/ประเภทขึ้นไป
- Multinomial DV หมายถึงชื่อของตัวแปรตามที่ใช้
- BASE=LAST หมายถึงประเภทหรือกลุ่มที่ใช้เป็น reference categoryคือกลุ่มสุดท้าย ในที่นี้นักสถิติอาจใช้กลุ่มแรกหรือกลุ่มอื่นเป็น reference category ก็ได้ ค่า parameter ที่ได้ในกลุ่มอื่นจะถูกใช้เปรียบเทียบกับค่าของกลุ่มที่เป็น reference category
- ORDER=ASCENDING หมายถึงตัวเลขแทนลำดับที่กลุ่มจะเริ่มจากน้อยไปหามาก
- Factor หมายถึงตัวแปรอิสระที่เป็นนามบัญญัติ
- Covariate หมายถึงตัวแปรที่มีค่าต่อเนื่อง
- MODEL=Factor Covariate หมายถึงแบบจำลองที่ใช้ตัวแปรอิสระที่เป็นนามบัญญัติ และตัวแปรอิสระที่มีค่าต่อเนื่องในการพยากรณ์ตัวแปรตาม
- INTERCEPT=INCLUDE หมายถึงแบบจำลองที่มี INTERCEPT รวมอยู่ด้วย

- PRINT เป็นคำสั่งให้แสดงผล
- PARAMETER SUMMARY ตารางแสดงค่า parameter ทั้งหมด
- LRT หมายถึง Likelihood ratio test
- CPS หมายถึง ตาราง case processing summary
- MFI หมายถึงตาราง model fitting information

Output ที่ได้จากโปรแกรมสำเร็จรูป

1. Case processing summary table:

แจกแจงให้เห็นความถี่ของค่าตัวแปรตามที่เกิดขึ้นในแต่ละกรณี ตลอดจนสัดส่วนร้อยละของค่าตัวแปรตามในแต่ละกรณี และจำนวนของmissing data

2. Model fitting information table:

แสดงให้เห็นสถิติสามารถประเมินผลทางสถิติของintercept only model หรือnull model (ที่ไม่ได้ใช้ตัวแปรอิสระมาช่วยในการพยากรณ์) กับ full หรือ final model (ที่ใช้ตัวแปรอิสระมาช่วยในการพยากรณ์) โดยสามารถพิจารณาจากค่าดัชนีหลายอย่างดังนี้

- Akaike Information Criterion (AIC) เป็นดัชนีที่ใช้ในทฤษฎีข้อมูลซึ่งมีว่าหากจะให้แบบจำลองสามารถอธิบายข้อมูลได้ดีขึ้น อาจจำเป็นที่จะต้องเพิ่มตัว parameter ขึ้นมา ซึ่งทำให้แบบจำลองมีความซับซ้อนมากขึ้น ในการเลือกแบบจำลองโดยใช้AIC กำหนดให้เลือกแบบจำลองที่มี AIC ต่ำที่สุด จึงจะได้แบบจำลองที่ดีที่สุด

- Bayesian Information Criterion(BIC) เป็นดัชนีอีกอย่างที่คล้ายคลึงกับ AIC ในการเลือกแบบจำลองโดยใช้BIC กำหนดให้เลือกแบบจำลองที่มี BIC ต่ำที่สุด จึงจะได้แบบจำลองที่ดีที่สุด

ทั้ง AIC และ BIC อาจมีค่าติดลบได้ และแบบจำลองที่มีค่าทั้งสองติดลบเป็นแบบจำลองที่ดีกว่าแบบจำลองที่มี

ค่า AIC หรือ BIC เป็นบวก

- -2Log Likelihood(-2LL) แสดงความผันผวนในค่าของตัวแปรตามที่แบบจำลองไม่สามารถอธิบายได้ ในที่นี้

-2LL ของ full model จะลดลงต่ำกว่า -2LL ของ null model ยิ่งลดลงมากเท่าใด ก็ยิ่งแสดงให้เห็นความ fit ของ

แบบจำลองมากขึ้นเท่านั้น(ตัวแปรอิสระเข้ามาช่วยอธิบายความผันผวนในค่าของตัวแปรตาม เป็นผลให้ความผันผวน

ของตัวแปรตามที่ไม่สามารถอธิบายได้ลดลง)

•Likelihood ratio Chi-square test เป็นการทดสอบ goodness-of-fit อีกทางเลือกหนึ่งที่จะช่วยให้นักสถิติพิจารณาว่า การลดลงใน -2LL มีนัยสำคัญทางสถิติหรือไม่ อย่างไรก็ตามให้พึงระลึกอยู่เสมอว่า ค่าChi-square ที่ได้มีแนวโน้มที่จะมีนัยสำคัญหากขนาดของกลุ่มตัวอย่างที่ใช้เป็นกลุ่มตัวอย่างขนาดใหญ่

3. Warnings

ในบางครั้งจะมีข้อความเตือนดังตัวอย่างต่อไปนี้

There are X (Y%) cells (i.e., dependent variable levels by subpopulations)with zero frequencies

นั่นหมายความว่า มีเซลล์บางเซลล์ที่ค่าความถี่ที่ได้จากการประมาณการมีค่าปรากฏเป็นศูนย์ ซึ่งสื่อให้เห็นแนวว่า อาจเกิดปัญหาขึ้น อย่างไรก็ตามหากจำนวน cell หรือร้อยละของเซลล์มีจำนวนน้อย เราอาจเพิกเฉยค่าเตือนนี้ และข้ามไปสรุปผลจาก goodness-of-fit table (ดังจะได้กล่าวต่อไป) ได้โดยปลอดภัย แต่หากจำนวน cell หรือร้อยละของเซลล์มีจำนวนมาก เราจำเป็นต้องกังวล และหลีกเลี่ยงการตีความผลที่ได้จาก goodness-of-fit table

4. Goodness-of-fit table

เป็นตารางแสดงค่า Chi-square สองค่า ได้แก่ Pearson Chi-square และ Deviance Chi-square หากจำนวน cell หรือร้อยละของเซลล์ในส่วนที่เป็นค่าเตือนในข้อ 3 ก่อนหน้ามีจำนวน/สัดส่วนที่สูง เราต้องหลีกเลี่ยงไม่ตีความผลจากตารางนี้ แต่หากจำนวนเซลล์ หรือร้อยละของเซลล์ที่มีความถี่ที่ได้จากการประมาณการเท่ากับศูนย์มีจำนวน/สัดส่วนที่ต่ำหรือไม่มีเลย เราสามารถเข้าไปตีความผลที่ได้จากตารางนี้ได้เลย ซึ่งเป็นที่คาดไว้ว่า ทั้ง Pearson Chi-square และ Deviance Chi-square จะต้องมีนัยสำคัญทางสถิติ จึงจะสามารถสรุปได้ว่า แบบจำลองใช้ได้ผลและเป็นประโยชน์ในการพยากรณ์ และเราสามารถดำเนินการในขั้นตอนต่อไป

5. Pseudo-R-square table

เป็นตารางแสดงให้เห็นค่า R-square ของ Cox and Snell, Nagelkerke และ McFadden ซึ่งให้ข้อมูลคร่าว ๆ ถึง

สัดส่วนของความผันผวนในตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ หากค่า R-square ที่ได้มีค่าใกล้เคียงศูนย์ แสดงว่าแบบจำลองไม่ประสบความสำเร็จมากนักในการพยากรณ์ แต่หาก R-square ที่ได้มีค่าใกล้เคียงหนึ่ง แสดงว่าแบบจำลองทำงานได้ผล และสามารถใช้อธิบายความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระได้ อย่างไรก็ตามเนื่องจากเป็นค่าประมาณการ นักสถิติอาจไม่ให้ความสำคัญกับการตีความผลที่ได้จากตารางนี้มากนัก

6. Parameter Estimate table

เป็นตารางแสดงให้เห็นผลของตัวแปรอิสระแต่ละตัวว่ามีความสำคัญหรือไม่ ค่าสัมประสิทธิ์หน้าตัวแปรอิสระ (β 's)

ตลอดจนค่า odd-ratio หากค่าสัมประสิทธิ์เป็นบวก มีความหมายว่า ตัวแปรอิสระนั้นมีส่วนเพิ่มความเป็นไปได้ (likelihood) ในผลที่เป็นตัวแปรตาม แต่หากค่าสัมประสิทธิ์เป็นลบ มีความหมายว่า ตัวแปรอิสระนั้นมีส่วนที่จะลดความเป็นไปได้ (likelihood) ในผลที่เป็นตัวแปรตาม

7. Classification table

เป็นตารางที่แสดงการเปรียบเทียบผลของตัวแปรตามที่เกิดขึ้นจริงกับผลของตัวแปรตามที่พยากรณ์ขึ้นจากแบบจำลอง ยิ่งสัดส่วนการพยากรณ์มีความถูกต้องเท่าใด ก็ยิ่งแสดงให้เห็นความมีประสิทธิภาพของแบบจำลองที่ใช้ตัวแปรอิสระในการพยากรณ์ค่าของตัวแปรตามมากขึ้นเท่านั้น

เพื่อเป็นการป้องกันความผิดพลาด นักสถิติมักจะคำนวณหาสัดส่วนความถูกต้องในการพยากรณ์ขึ้นจากข้อมูลที่มีอยู่ หรือที่เรียกว่า สัดส่วนความถูกต้องในการพยากรณ์ที่ขึ้นกับโชค (by chance) โดยคำนวณหาผลรวมของสัดส่วนผลของตัวแปรตามที่เกิดขึ้นยกกำลังสองในแต่ละกรณี ตัวอย่างเช่น จากข้อมูลที่มีอยู่ ตัวแปรตามมีค่าแสดงได้ดังตารางต่อไปนี้

	ร้อยละ
A	30
B	50
C	20

สัดส่วนความถูกต้องในการพยากรณ์ที่ขึ้นกับไซคคำนวณได้ดังนี้

$$(.30)^2 + (.50)^2 + (.20)^2 = .38 \text{ หรือ } 38\%$$

ในทางปฏิบัติ นักสถิติต้องการผลสรุปเป็นเด็ดขาด ดังนั้นอาจให้ความถูกต้องในการพยากรณ์ที่เกิดจากไซคมี

ความแปรปรวนเพิ่มขึ้นได้ราว 25% เมื่อเป็นเช่นนี้เกณฑ์ที่นักสถิติใช้ในการคำนวณหาความถูกต้องในการพยากรณ์ที่เกิดจากไซค=47.5% (38X1.25)

หาก classification table แสดงความถูกต้องในการพยากรณ์ไม่เกิน 47.5% เราไม่อาจวางใจในการใช้แบบจำลอง MLR ในการใช้ตัวแปรอิสระอธิบายตัวแปรตามได้ แต่หากความถูกต้องในการพยากรณ์ตาม classification table สูงและมีค่าห่างจาก 47.5 % มาก เมื่อนั้นแบบจำลองนี้มีประสิทธิภาพและสามารถเอาไปใช้ในการพยากรณ์ตัวแปรตามจากตัวแปรอิสระได้

Contribution this issue: ดร. ดนัย ปัตตพงษ์

งานวิจัยที่อ้างอิงบทความวิชาการนี้

ณัฐพัชร์ อภิรุ่งเรืองสกุลและคณะ “ การวิเคราะห์ปัจจัยที่มีผลต่อการทำธุรกรรมทางการเงินผ่าน QR Code ในกรุงเทพมหานคร “ วารสารปัญญาภิวัฒน์ ฉบับพิเศษ ปีที่ 10.2018.

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 ซอฟต์แวร์ en@gex@cel® สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาาระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)