Ordered Logistic Regression(OLOGIT)

Ordered logistic regression หรือ ordinal regression หรือ cumulative logit model จะเป็นเทคนิคทางสถิติที่ใช้ตัวแปรอิสระพยากรณ์ค่าของตัวแปรตาม โดยตัวแปรตามเป็นตัวแปรที่มีค่าแสดงระดับ หากแต่เราไม่สามารถกำหนดระยะห่างที่แท้จริงระหว่างระดับได้ ในบทความนี้เราจะมุ่งศึกษาถึงกรณีที่มีการใช้ logit link function เท่านั้น ในกรณีนี้ค่า parameter หรือค่าสัมประสิทธิ์ที่อยู่หน้าตัวแปรอิสระจะไม่ขึ้นกับระดับของตัวแปรตาม ในบางครั้งแบบจำลองนี้จึงมีชื่อเรียกว่า proportional odds model หรือparallel lines modelหรือparallel regressions model

OLOGIT บางทีเรียกว่า cumulative logit model ที่เป็นเช่นนี้ก็เพราะ สมการregression ที่ใช้จะอยู่ในรูป

$$\ln (F_{ij} / 1-F_{ij}) = \beta_{Oj} - (\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

โดยที่ F_{ij} เป็นค่าความถี่สะสม กล่าวคือ

F_{i1} คือค่า Prob(Y=1) หรือค่าความถี่ของตัวแปรตามที่อยู่ในลำดับที่ 1(ลำดับต่ำที่สุด)

F_{i2} คือค่า Prob(Y$\leq$ 2) หรือค่าความถี่สะสมของตัวแปรตามที่มีค่าอยู่ในลำดับที่ 2 และ 1

F_{i3} คือค่า Prob(Y$\leq$ 3) หรือค่าความถี่สะสมของตัวแปรตามที่มีค่าอยู่ในลำดับที่ 3 , 2 และ 1 เป็นต้น

อนึ่ง ต้องทำความเข้าใจว่า โปรแกรมสำเร็จรูปอื่นๆ (Stata,SAS เป็นต้น) ที่ใช้ใน OLOGIT มีข้อกำหนดเกี่ยวกับตัวแปรตามและตัวแปรต้นต่างกัน สำหรับโปรแกรมสำเร็จรูปนี้ตัวแปรตามเรียงลำดับจากน้อยไปหามาก (ascending) ในขณะที่ตัวแปรต้นเรียงจากมากไปหาน้อย (descending) เมื่อเป็นเช่นนี้ odds ratio (exp(β)) ในกรณีที่ค่า β เป็นบวก จะแสดง odd ที่ตัวแปรตามจะขยับลำดับที่สูงขึ้น หากค่าตัวแปรต้น (X) เพิ่มขึ้นหนึ่งหน่วย ในโปรแกรมสำเร็จรูปบางโปรแกรม การเพิ่มขึ้นของตัวแปรต้น (X) หนึ่งหน่วยจะเพิ่ม odd ที่ตัวแปรตามจะขยับลำดับที่ต่ำลง

จะใช้เทคนิคนี้ในสถานการณ์เช่นใด?

หากเรามีตัวแปรตามที่มีค่าแสดงระดับ OLOGIT จะมี statistical power(ความน่าเชื่อถือในการพยากรณ์) สูงกว่า MLR(Multinomial Logistic Regression)

ตัวอย่างของการใช้เทคนิคนี้

วงการเครือข่ายร้านอาหารจานด่วน: บริษัทวิจัยตลาดแห่งหนึ่งต้องการศึกษาดูว่ามีปัจจัยอะไรที่เป็นตัวกำหนดขนาดของโซดาที่ลูกค้าสั่งซื้อ โดยตัวแปรอิสระที่ใช้คือ ประเภทของแซนด์วิชที่ลูกค้าสั่งมาทาน (แฮมเบอร์เกอร์ / เบอร์เกอร์ไก่), มันฝรั่งทอด (สั่ง / ไม่สั่ง) อายุของผู้บริโภค ส่วนตัวแปรตามได้แก่ขนาดของโซดา (เล็ก กลาง ใหญ่ ใหญ่พิเศษ)

วงการศึกษา : นักวิจัยทำการศึกษาปัจจัยที่มีผลต่อการตัดสินใจของนักศึกษาปริญญาตรีในการเรียนต่อระดับปริญญาโท ตัวแปรตามเป็นระดับความคิดเห็นของนักศึกษา ได้แก่ ไม่มีทางเป็นไปได้ มีความเป็นไปได้บ้าง มีความเป็นไปได้มาก ส่วนตัวแปรอิสระได้แก่ สถานะการศึกษาของบิดามารดา ประเภทของสถาบันที่เรียนอยู่ (ของรัฐ / ของเอกชน) และเกรดโดยเฉลี่ย

วงการสำรวจอวกาศ : นักสถิติได้ใช้ OLOGIT ในการพยากรณ์ความเสี่ยงของการปล่อยกระสวยอวกาศ โดยใช้ตัวแปรต้นที่สำคัญสามตัว ได้แก่ 1.จำนวนครั้งที่ไอร้อนจากจรวดช่วยขับเคลื่อน (booster) อาจทำลายจุดเชื่อมระหว่างถังเชื้อเพลิงภายนอก (external propellant tank) และจรวดช่วยขับเคลื่อน 2.อุณหภูมิวันที่ปล่อยยาน (หากอุณหภูมิลดลงต่ำมาก จะทำให้ยางรัดจุดเชื่อม(rubber O-ring) ระหว่างถังเชื้อเพลิงและจรวดช่วยขับเคลื่อนแข็งตัวและไม่ยืดหยุ่น มีความเสี่ยงกับการที่ยางรัดจะล้มเหลวในระหว่างใช้งาน 3.จำนวนวัน (date) นับจากวันที่ 1 มกราคม 1960 (วันเริ่มต้นโครงการ) ไปจนถึงวันปล่อยกระสวยอวกาศ หากจำนวนวันยิ่งมาก อาจบ่งบอกว่า อุปกรณ์หรือฮาร์ดแวร์บางอย่างที่ผลิตขึ้นและเก็บไว้ในสต๊อก อาจเสื่อมสภาพไปตามกาลเวลา นักสถิติพบว่า OLOGIT สามารถพยากรณ์ความเป็นไปได้ว่าจะเกิดเหตุการณ์ที่ยางรัดจุดเชื่อมจะล้มเหลวด้วยความน่าจะเป็นสูงถึง 0.99959 นักสถิติยังเสริมด้วยว่า หากเจ้าหน้าที่ขององค์การนาซ่าได้เห็นข้อมูลเตือนภัยเหล่านี้ ก็จะยกเลิกภารกิจส่งกระสวยอวกาศ Challenger ที่กลายเป็นโศกนาฏกรรมเมื่อวันที่ 28 มกราคม พ.ศ. 2529 ใน

วงการโภชนาการ : นักโภชนาการสนใจศึกษาปัญหาการขาดสารอาหารของเด็กในบังคลาเทศ ตัวแปรอิสระที่ศึกษาได้แก่ อายุของเด็ก ช่วงที่เด็กเกิดต่อเนื่องกันมีระยะเวลาสั้นเกินไป การศึกษาของมารดา การมีอาหารเพียงพอช่วงตั้งครรภ์ สถานะครอบครัว ดัชนีพฤติกรรมการให้อาหารแก่เด็ก ภาวะการมีไข้ตัวร้อน การมีโรคติดเชื้อทางเดินหายใจ การมีโรคท้องร่วง ส่วนตัวแปรตามได้แก่ ภาวะขาดสารอาหารอย่างรุนแรง ภาวะขาดสารอาหารปานกลาง ไม่มีภาวะขาดสารอาหาร

วงการผลิตรถยนต์ : นักสถิติอาจสนใจสถิติประวัติของรถยนต์ที่มีการซ่อมในระยะใช้งานห้าปีแรกเพื่อประเมินคุณภาพของรถยนต์ที่ถูกผลิตขึ้นมา โดยตัวแปรตามอาจแบ่งออกเป็นรถที่มีประวัติต้องซ่อมบ่อย(poor) รถที่มีประวัติการซ่อมพอประมาณ (fair) รถที่มีประวัติการซ่อมน้อยมาก(good) รถที่ไม่มีประวัติการซ่อมเลย(excellent) ส่วนตัวแปรอิสระได้แก่ ชื่อผู้ผลิต ขนาดความยาวของรถ อัตราการบริโภคน้ำมันเชื้อเพลิง (mpg.)

วงการแพทย์ : ในช่วงที่มีการระบาดของ Covid-19 ในประเทศจีน ได้มีการใช้ OLOGIT ไปใช้ในการวินิจฉัยปัจจัยที่มีผลทำให้อาการป่วยของคนไข้รุนแรง ตัวแปรตามได้แก่ระดับความรุนแรง แบ่งออกเป็น รุนแรงน้อย รุนแรงปานกลาง รุนแรงมาก รุนแรงถึงขั้นวิกฤติ ตัวแปรอิสระได้แก่ underlying conditions, อายุ ดัชนีชี้วัดจากห้องแลป (โปรตีน D-dimer โปรตีน CRP เอนไซม์ LDH โปรตีน troponin I) เซลล์เม็ดเลือดขาว การฉีดวัคซีน COVID-19 และการบำบัดรักษาด้วยยาแอนตี้ไวรัส

วงการดูแลสุขภาพ :โรงพยาบาลต้องการศึกษาความสัมพันธ์ระหว่างระดับความอ้วน (ปกติ / น้ำหนักมากเกิน/ อ้วน) กับตัวแปรอิสระซึ่งประกอบด้วยเพศ (ชาย/หญิง) อายุ ระดับกิจกรรมที่ทำในระหว่างสัปดาห์ซึ่งรวมถึงการออกกำลังกาย

วงการกีฬา: นักวิทยาศาสตร์ด้านการกีฬาต้องการศึกษาจำนวนเหรียญรางวัล (ทอง/เงิน/บรอนซ์) ที่ได้จากกีฬาว่ายน้ำ โดยอาศัยตัวแปรอิสระที่ประกอบด้วยชั่วโมงที่ใช้ในการฝึก การรับประทานอาหาร อายุ และความนิยมชมชอบของประชากรในท้องถิ่นเกี่ยวกับกีฬาว่ายน้ำนี้

สมมติฐานที่จำเป็นในการใช้เทคนิคทางสถิตินี้

สมมติฐานที่ 1: ตัวแปรตามต้องมีมาตรวัดเป็นระดับ เช่น มาตรวัดตามแนวลิเกิรท์ 7ระดับ (จากstrongly disagree -ไม่เห็นด้วยอย่างยิ่ง , moderately disagree-ไม่เห็นด้วยพอประมาณ , disagree-ไม่เห็นด้วย , neutral-มีความรู้สึกกลางๆ ,agree-เห็นด้วย , moderately agree-เห็นด้วยพอประมาณ , strongly agree-เห็นด้วยอย่างยิ่ง) มาตรวัดความพึงพอใจสามระดับ (not very much-ชอบไม่มากนัก ,ok-ชอบปานกลาง , a lot-ชอบมาก)

สมมติฐานที่ 2: มีตัวแปรอิสระหนึ่งหรือมากกว่าที่เป็นตัวแปรที่มีค่าต่อเนื่อง (continuous variable) ตัวแปรที่มีค่าแบ่งเป็นระดับ (ordinal variable) ตัวแปรที่เป็นนามบัญญัติ (categorical variable ซึ่งรวมถึงตัวแปรที่มีเพียงสองประเภท)ตัวอย่างของตัวแปรเหล่านี้เช่น

ตัวแปรที่มีค่าต่อเนื่อง: อายุ (ปี) น้ำหนัก (กิโล) รายได้ (จำนวน) สติปัญญา (IQ score) ผลการสอบ (คะแนน 0 ถึง 100)

ตัวแปรที่มีค่าแบ่งเป็นระดับ :ขนาดของสินค้า (เล็ก กลาง ใหญ่ จัมโบ้) ความพึงพอใจ (ต่ำ กลาง สูง) สถานภาพทางสังคม (ต่ำ กลาง สูง) ระดับการศึกษา (อนุปริญญาหรือต่ำกว่า ปริญญาตรี ปริญญาโท ปริญญาเอก)

ตัวแปรที่เป็นนามบัญญัติ:เพศ(ชาย/หญิง) ศาสนา (พุทธ/คริสต์/ฮินดู/มุสลิม/อื่นๆ) เชื้อชาติ (ผิวขาว/ผิวดำ/เอเชีย/ฮิสปานิค/อื่นๆ)

สมมติฐานที่ 3 : ต้องไม่มี multicollinearityในระหว่างตัวแปรอิสระ

สมมติฐานที่ 4: ตัวแปรอิสระแต่ละตัวต้องมีผลต่อตัวแปรตามเหมือนกันในทุกระดับของตัวแปรตาม ที่เรียกว่า proportional odds assumption หรืออธิบายได้ง่ายๆในกรณีของภาพ 2 มิติ หรือ 3 มิติ ก็คือ slope ของ logit functionsที่ใช้ในการพยากรณ์ตัวแปรตามแต่ละระดับจะขนานกันแม้จะมี intercept ต่างกันตามระดับค่าของตัวแปรตาม โดยlogit functions จะไม่มีการตัดข้ามกัน วิธีที่ทำให้นักสถิติมั่นใจว่าไม่มีการละเมิด proportional odds assumption ในแบบจำลองที่ศึกษาอยู่ ก็คือต้องมีการทดสอบโดยใช้ test of parallel lines ซึ่ง null hypothesis ระบุว่า slope ของ logit functions เหมือนกัน หากเรายอมรับnull hypothesis

หมายความว่า เราสอบผ่านสมมติฐานข้อนี้ และสามารถดำเนินการวิเคราะห์ต่อไปโดยใช้ OLOGIT ได้ แต่หากปฏิเสธ null hypothesis เราต้องหันไปใช้MLR มาวิเคราะห์แทน อย่างไรก็ตามการใช้ MLR ทำให้เราสูญเสียประสิทธิภาพทางสถิติไป เนื่องจากเราจงใจละทิ้งข้อมูลที่บ่งบอกลำดับที่ของตัวแปรตาม และไม่ได้ใช้ข้อมูลนั้น ๆทั้ง ๆที่เรามีข้อมูลอยู่แล้ว ทำให้เราอาจต้องมีการประมาณค่าparameter เพิ่มขึ้นกว่าที่จำเป็น ลดโอกาสที่จะทำให้แบบจำลองที่ใช้ MLR มีนัยสำคัญทางสถิติ

การใช้โปรแกรมสำเร็จรูปมาใช้ในการแก้ปัญหาที่เกี่ยวกับ ordered logistic regression

สามารถทำได้สามวิธีดังนี้

1. ใช้ menu ของโปรแกรมสำเร็จรูปที่เรียกว่า drop-down menu ในการวิเคราะห์

   ในการกำหนดรูปแบบ model เราอาจเลือกใช้ default model ซึ่งจะรวม intercept ตัวแปรอิสระที่เป็นนามบัญญัติ (factors) ตัวแปรอิสระที่มีค่าต่อเนื่อง (covariates) หรืออาจเลือกใช้ location model ในกรณีที่ต้องการดูผลหลัก (main effects) และผลที่เกิดจาก interaction effects ระหว่างตัวแปรอิสระสองตัวหรือมากกว่า

2.ใช้ PLUM (polytomous universal model) syntax command ซึ่งมีรูปแบบดังนี้

```
PLUM  OrdinalDV  BY Factor WITH Covariate
/LINK=LOGIT
/PRINT= FIT  PARAMETER SUMMARY  TPARALLEL.
```

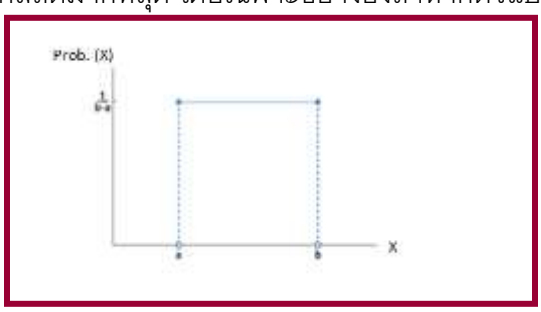
คำอธิบาย

- PLUM เป็น subcommandของโปรแกรมสำเร็จรูปSPSS ใช้ในการวิเคราะห์ปัญหาที่เป็น Ordered logistic regression
- Ordinal DVหมายถึงตัวแปรตามที่มีค่าแบ่งเป็นระดับ

●Factor หมายถึงตัวแปรอิสระที่เป็นนามบัญญัติ

●Covariate หมายถึงตัวแปรอิสระที่มีค่าต่อเนื่อง

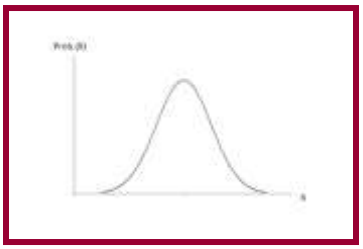
●LINK=LOGIT หมายถึงเลือกใช้ logitเป็น link function (นักสถิติอาจเลือกใช้ link functionอื่นๆ เช่น CAUCHITหรือ CLOGLOG หรือ NLOGLOG หรือ PROBIT) แต่ logit function เป็น link function ที่ได้รับความนิยมจากนักสถิติมากที่สุด โดยเฉพาะอย่างยิ่งถ้าหากตัวแปรตามมีการกระจายที่เป็น uniform distribution.



รูปภาพแสดง
Uniform distribution

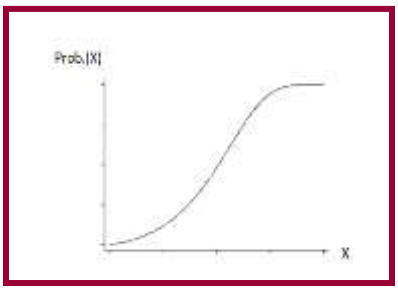
หากเราตัดสินใจใช้ link function อื่นแทนที่จะเป็น logit link function เราจะไม่สามารถคำนวณหาค่าของ odds ratios ได้เหมือนกรณีที่ใช้ logit link function ยิ่งไปกว่านั้นการตีความค่าสถิติต่างๆจากตาราง parameter estimates ไม่สามารถทำได้เหมือนกับกรณีที่ใช้ logit link function

- นักสถิติอาจเลือกใช้ probit เป็นlink functionเมื่อใดก็ตามที่ตัวแปรตามมีการกระจายแบบ normal



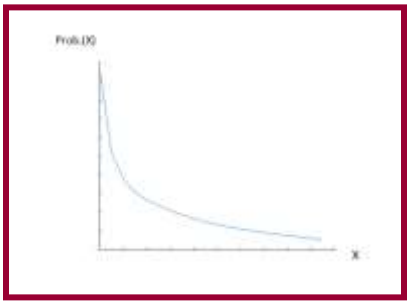
รูปภาพแสดง
Normal distribution

-เมื่อใดก็ตามที่โอกาสที่ตัวแปรตามจะมีค่าในลำดับสูงมากกว่าค่าในลำดับต่ำ นักสถิติจะเลือกใช้ complementary log-log เป็น link function หรือที่รู้จักกันว่า continuation ratio model/proportional hazard model



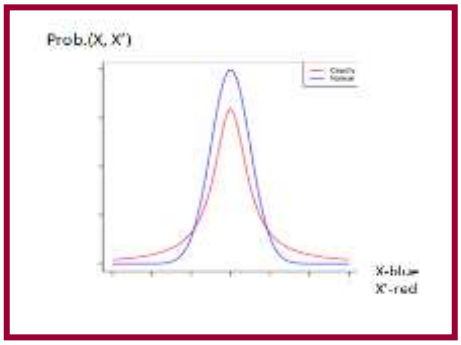
รูปภาพแสดง
Log-log function

-ในทางกลับกันเมื่อใดก็ตามที่โอกาสที่ตัวแปรตามจะมีค่าในลำดับต่ำมากกว่าค่าในลำดับสูง นักสถิติจะเลือกใช้ negative log-logเป็น link function



รูปภาพแสดง
Negative log-log function

-และเมื่อใดก็ตามที่ตัวแปรตามมีค่ากระจุกอยู่ตรงค่าต่ำ และค่าสูง นักสถิติจะเลือกใช้ Cauchit เป็นlink function ภาพด้านล่างแสดงการเปรียบเทียบการกระจายของ X ที่เป็น Normal และ X' ที่เป็น Cauchy



รูปภาพแสดงการกระจาย
ของ Normal vs. Cauchy

- PRINT เป็นคำสั่งให้แสดงผล
- FITหมายถึง model fit statistics
- PARAMETER หมายถึงตาราง parameter estimates
- SUMMARY หมายถึง summary statistics
- TPARALLEL หมายถึงการทดสอบสมมติฐาน proportional odds assumptionเพื่อตรวจสอบว่า slope ของ logit functions ที่ใช้ในการพยากรณ์ตัวแปรตามแต่ละระดับจะขนานกันแม้จะมี intercept ต่างกันตามระดับค่าของตัวแปรตาม(ดูสมมติฐานที่ 4 ก่อนหน้านี้)

3.ใช้ menu ของโปรแกรมสำเร็จรูปในส่วนที่เป็น generalized linear model (GZLM) ทั้งนี้เนื่องจาก OLOGIT เป็นกรณีพิเศษของ GZLM

Output ที่ได้จากโปรแกรมสำเร็จรูป

1. Warnings เป็นการเตือนให้นักสถิติทราบว่า สัดส่วนของ cell ที่เกิดจากการไขว้ข้อมูล(cross tabulation) ระหว่างระดับของตัวแปรตามกับตัวแปรอิสระที่มีค่าความถี่ที่คาดว่าจะเป็น (expected frequency)ต่ำกว่า 5 มีเป็นร้อยละเท่าใด หากเป็นสัดส่วนที่สูง เกิน 20% จะทำให้ความน่าเชื่อถือของผลทางสถิติที่จะได้ตามมาลดน้อยถอยลง อย่างไรก็ตามคำเตือนนี้อาจแสดงสัดส่วน cell ที่มีค่าความถี่ที่คาดว่าจะเป็นต่ำกว่า 5 สูงเกินความเป็นจริง ยิ่งโดยเฉพาะอย่างยิ่งเมื่อรวม covariate space เข้าไปด้วย เพื่อความมั่นใจนักสถิติจึงต้องพิจารณาเฉพาะ factor space ซึ่งก็คือจำกัดการไขว้ข้อมูล(cross tabulation) ระหว่างระดับของตัวแปรตามกับตัวแปรอิสระที่เป็นนามบัญญัติหรือที่เรียกว่า factor เท่านั้น หากมี cell ที่มีค่าความถี่เป็น 0 หรือน้อยกว่า 5 ผลทางสถิติที่ตามมาจะมีความเชื่อถือน้อยลงไป ประเด็นที่ cell มีค่าความถี่น้อยไปเรียกว่าปัญหาความเพียงพอของความถี่ในแต่ละเซลล์ (cell count adequacy) ซึ่งจะมีการขยายความเพิ่มเติมต่อไป

2.Case processing summary แสดงการแจกแจงค่าความถี่และร้อยละของตัวแปรตามและตัวแปรอิสระที่เป็นนามบัญญัติที่มักจะเรียกกันว่าปัจจัย (factor)

3.Model Fitting Information table ให้ข้อมูลเกี่ยวกับความผันผวนในตัวแปรตามที่ไม่สามารถอธิบายได้แม้จะมีการนำตัวแปรอิสระเข้ามาในแบบจำลอง (ที่เรียกว่า full model/final model ซึ่งจะมีค่า -2log likelihood ที่ต่ำกว่า) เปรียบเทียบกับความผันผวนในตัวแปรตามที่ไม่สามารถอธิบายได้จากแบบจำลองที่ไม่อาศัยตัวแปรอิสระใดๆเลย (ที่เรียกว่า null model/intercept only model ซึ่งจะมีค่า -2 log likelihood ที่สูงกว่า) ผลต่างระหว่างความผันผวนทั้งสองจะถูกนำมาใช้ในการทดสอบสมมติฐานว่า ค่าของ regression coefficients หน้าตัวแปรอิสระทุกตัวมีค่าเป็นศูนย์ (0)หรือไม่ หากการทดสอบโดย Likelihood ratio Chi-square ไม่แสดงความมีนัยสำคัญ หมายความว่า regression coefficients หน้าตัวแปรอิสระเป็นศูนย์ และตัวแปรอิสระไม่มีบทบาทในการอธิบายความผันผวนในตัวแปรตามได้เลย แต่หากการทดสอบโดย Likelihood ratio Chi-square แสดงความมีนัยสำคัญ หมายความว่า อย่างน้อยก็มีตัวแปรอิสระหนึ่งตัวที่มี regression coefficient ที่ไม่ใช่ศูนย์ และการนำตัวแปรอิสระเข้ามาในแบบจำลองมีส่วนช่วยลดทอนความผันผวนในตัวแปรตามส่วนที่ไม่สามารถอธิบายได้ลง อย่างไรก็ตามขอทำความเข้าใจว่า การนำตัวแปรอิสระเข้ามามิได้หมายความว่าจะช่วยลดทอนความผันผวนในตัวแปรตามที่ไม่สามารถอธิบายลงถึงระดับที่มีนัยสำคัญทุกกรณี

4. **Goodness-of-fit table ให้ข้อมูลเกี่ยวกับการทดสอบสมมติฐาน H_O : การใช้ OLOGIT สามารถใช้อธิบายข้อมูลได้ดี (The fit is good)** โดยทำการเปรียบเทียบระหว่างค่าความถี่ที่ได้จากการประมาณการ (expected frequency) และ ค่าความถี่ที่ได้จากการสังเกต (observed frequency)ของแต่ละ cell หากแบบจำลองที่พิจารณาอยู่มีความถูกต้อง ค่าความถี่ทั้งสองจะมีจำนวนใกล้เคียงกัน พิจารณาจาก Pearson's หรือ Deviance Chi-square ที่ไม่มีนัยสำคัญ อย่างไรก็ตามหากขนาดของกลุ่มตัวอย่างมีขนาดใหญ่ ความแตกต่างเพียงเล็กน้อยระหว่างค่าความถี่ที่ได้จากการประมาณการกับค่าความถี่ที่ได้จากการสังเกตอาจมีผลทำให้ผลของการทดสอบที่แสดงด้วย goodness-of-fit tableมีนัยสำคัญ หากเป็นเช่นนั้น นักสถิติต้องลดขนาดของกลุ่มตัวอย่างลงและลองทำการทดสอบซ้ำ

อนึ่งการทดสอบ goodness-of-fit นี้มักเชื่อถือไม่ค่อยได้หากความถี่ในแต่ละเซลล์ มีไม่เพียงพอ โดยเฉพาะอย่างยิ่งหาก มีcovariateในmodel หรือขนาดของกลุ่มตัวอย่างมีขนาดเล็ก

5.**Pseudo R-square table** เป็นตารางแสดงค่าของ Pseudo R-square (แปลว่า R-square เทียม)เพื่อให้นักสถิติได้มีข้อมูลในลักษณะใกล้เคียงกับ R-square ที่ได้จากการวิเคราะห์การถดถอย(regression analysis) โดยจะประกอบด้วย Nagelkerke's R-square ,Cox&Snell R-square(ที่ค่าสูงสุดไม่มีวันถึง 1) และ McFadden's R-square โดย Nagelkerke's R-squareจะเป็นค่า R-square เทียมที่นิยมถูกหยิบยกขึ้นมารายงานในงานวิจัยมากที่สุด อย่างไรก็ตาม R-square เทียมที่คำนวณขึ้นมานี้เป็นเพียงดัชนีหยาบๆไม่สามารถตีความได้เฉกเช่น R-square ของการวิเคราะห์การถดถอย(OLS) ที่บ่งบอกสัดส่วนความผันผวนของตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระ และบรรดานักสถิติมักจะใช้ความระมัดระวัง โดยจะไม่จริงจังในการตีความจาก R-square เทียมเหล่านี้มากนัก

6. ตาราง **parameter estimates** แสดงค่าต่าง ๆดังนี้

- threshold จะเป็นค่าคงที่ซึ่งแบ่งแยกระดับของตัวแปรตามหนึ่งออกจากระดับอื่น โดยจำนวน threshold ที่แสดงจะมีจำนวนเท่ากันกับระดับของตัวแปรตามลบด้วยหนึ่งเสมอ

มีความเป็นไปได้ที่ threshold อาจไม่มีนัยสำคัญทางสถิติ หากเป็นเช่นนั้นอาจจำเป็นต้องมีการรวมระดับของตัวแปรตามบางระดับเข้าด้วยกัน อย่างไรก็ตามการมี threshold ที่ไม่มีนัยสำคัญทางสถิติมิได้หมายความว่า ตัวแปรอิสระที่มีอยู่จะไม่มีนัยสำคัญทางสถิติและแบบจำลองที่มีอยู่จะใช้ไม่ได้

-Location แสดงชื่อตัวแปรอิสระทั้งที่เป็นนามบัญญัติ(ที่เรียกว่า factor) และตัวแปรอิสระที่มีค่าต่อเนื่อง (ที่เรียกว่า covariates) ซึ่งอาจมีหรือไม่มีนัยสำคัญทางสถิติ

เปรียบเทียบการใช้ OLOGIT กับ MLR เมื่อมีตัวแปรตามที่มีการแบ่งค่าออกเป็นระดับ

ในกรณีที่ตัวแปรตามมีค่าแบ่งเป็นระดับ ผู้วิจัยอาจเลือกใช้ MLR (ดู Statistics Talks #30-32) ในการหาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม แต่ผลที่ตามมาคือประสิทธิภาพทางสถิติจะลดลง ทั้งนี้เนื่องจาก MLR จงใจไม่ใช้ประโยชน์จากข้อมูลของตัวแปรตามที่มีการแบ่งค่าออกเป็นระดับทั้ง ๆที่มีอยู่แล้ว และอาจทำให้ผู้วิจัยอาจจำเป็นต้องคำนวณหาค่า parameter เกินความจำเป็น ในทางสถิติเมื่อใดก็ตามที่ตัวแปรตามมีค่าแบ่งเป็นระดับ OLOGIT จะมี statistical power สูงกว่า MLR

Parallel lines test

Parallel lines test เป็นการทดสอบสมมติฐาน H_O : ตัวแปรอิสระต้องมี slope หรือ regression coefficients เท่ากัน ในทุกระดับของตัวแปรตาม (location parameters/slope coefficients are the same across response categories) หากพบว่าผลที่ได้จาก parallel lines test ไม่มีนัยสำคัญทางสถิติ นั่นหมายความว่า ไม่มีการละเมิดสมมติฐานข้อที่ 4 ของการใช้ OLOGIT และผลทางสถิติที่ได้จะมีความถูกต้องน่าเชื่อถือ

ในกรณีที่พบว่าผลของ parallel lines test มีนัยสำคัญทางสถิติ นักสถิติมีทางออกให้เลือกหลายทาง

1.ใช้ MLR แทน ซึ่งเป็นที่แน่นอนว่าจะทำให้เสีย power ทางสถิติไปเนื่องจากมีการตัดความสำคัญของข้อมูลแบ่งระดับออกไป

2.ตัดตัวแปรบางตัวออกหรือกำหนด model ขึ้นมาใหม่

3.เพิกเฉยผลที่เกิดจาก parallel lines test หากขนาดของกลุ่มตัวอย่างมีขนาดใหญ่ เพราะความแตกต่างเพียงเล็กน้อยใน slope มีผลทำให้การทดสอบแตกต่างอย่างมีนัยสำคัญ และเมื่อพิจารณาแล้ว เส้น regression ไม่ได้มีการไขว้ข้ามหรือตัดกัน

4.ลดกลุ่มตัวอย่างลงเหลือราว 200 แล้วทำการ re-run โปรแกรมขึ้นใหม่

5.ใช้ link function อื่นแทน logit link function

6.ยุบประเภทของตัวแปรอิสระให้มีจำนวนน้อยลง

7. ยุบประเภทของตัวแปรตามให้มีจำนวนน้อยลง

8.ตัดตัวแปรอิสระที่ไม่สำคัญบางตัวออก

9.นักสถิติบางท่านอาจนำเอา OLS (Ordinary Least Squares) มาใช้กับกรณีที่ตัวแปรตามมีค่าแบ่งออกเป็นระดับ โดยเฉพาะอย่างยิ่งถ้าตัวแปรมีหลายระดับ (5 หรือ 7 ระดับ)

ความน่าเชื่อถือของ goodness-of-fit table และจำนวนของ cell count ที่ต้องมีเพียงพอ

เมื่อใดก็ตามที่เซลล์ แต่ละเซลล์ ใน factor space ที่เกิดจากการสร้างตารางไขว้ (cross tabulation)ระหว่างตัวแปรตามกับตัวแปรอิสระที่เป็นตัวแปรนามบัญญัติมีค่าความถี่ที่คาดว่าจะเป็น (expected frequency) มีค่ามากกว่า 5 เมื่อนั้นผลการทดสอบใน ตารางgoodness-of-fit จะมีความน่าเชื่อถือ และนักสถิติสามารถสรุปผลการทดสอบทางสถิติได้จากตาราง goodness-of-fit ได้โดยตรง โดย model ที่ดี goodness-of-fit จะต้องไม่มีนัยสำคัญทางสถิติ (ความแตกต่างระหว่างค่าความถี่ที่ได้จากการสังเกตและค่าความถี่ที่ได้จากแบบจำลองในทุก ๆเซลล์มีค่าไม่แตกต่างกันมากอย่างมีนัยสำคัญ)

นักสถิติสามารถตรวจสอบ cell count adequacy ได้จาก factor space ที่เกิดจากตารางไขว้ระหว่างตัวแปรตามกับตัวแปรอิสระที่เป็นนามบัญญัติเท่านั้น โดยไม่นับรวม covariate space ที่เกิดจากตารางไขว้ระหว่างตัวแปรตามกับตัวแปรอิสระที่มีค่าต่อเนื่อง (covariates)

อนึ่ง เป็นที่เข้าใจว่า วิธีการทางสถิติที่เป็น logit หรือ probit regression อาศัยการคำนวณหา solution โดยใช้ maximum likelihood method(MLE) ซึ่งเป็นการคำนวณแบบกลับไปกลับมา จึงทำให้ต้องใช้ขนาดของกลุ่มตัวอย่างสูงมากกว่าการวิเคราะห์การถดถอยธรรมดา (ordinary least squares regression)

หากประสบปัญหาเรื่อง cell count adequacy นักสถิติมีทางเลือก 3 ทาง

1.เพิ่มขนาดของกลุ่มตัวอย่าง

2.มีการลงรหัสตัวแปรตามแบบมีระดับเสียใหม่ และ/หรือมีการลงรหัสตัวแปรอิสระที่เป็นนาม บัญญัติเสียใหม่เพื่อลดจำนวนตัวแปรลง

3.ตัดตัวแปรอิสระที่เป็นนามบัญญัติหนึ่งตัวหรือมากกว่านั้นออกจากแบบจำลองที่ใช้

การวิเคราะห์ต่อยอด : นักสถิติอาจพิจารณาตัดตัวแปรอิสระที่ไม่มีนัยสำคัญออกจากแบบจำลองทีละตัว จนกระทั่งcovariates ทุกตัวมีนัยสำคัญและอย่างน้อยก็มีตัวแปรอิสระที่เป็น factor หนึ่งระดับ ที่มีนัยสำคัญ

Contribution this issue : ดร. ดนัย ปัตตพงศ์