(Binary) Probit Regression

เป็นเทคนิคทางสถิติที่ใช้ในการพยากรณ์ค่าของตัวแปรตาม (dependent variable)ซึ่งมีค่าได้เพียง 2 ค่า โดยใช้ค่าของตัวแปรอิสระ คล้ายคลึงกับเทคนิค (Binary) logistic regression(ดู Statistics Talks# 16-19) โดยทั้ง logistic และ probit regression สามารถใช้ได้กับกรณีตัวอย่างต่อไปนี้

วงการเกษตรกรรม: นักวิจัยต้องการศึกษาดูว่า เกษตรกรเลือกปลูกพืชโดยใช้ปุ๋ยอินทรีย์หรือปุ๋ยเคมี (ตัวแปรตาม =1 ถ้าใช้ปุ๋ยอินทรีย์ =0 หากใช้ปุ๋ยเคมี) โดยพิจารณาจากชนิดของพืชที่ปลูก ขนาดของพื้นที่ ๆทำการเพาะปลูก ทำเลที่ตั้งของพื้นที่ ตลอดจนเข้าข่ายอยู่ในพื้นที่ชลประทานหรือต้องพึ่งพาฝน

วงการแรงงาน: ต้องการพิจารณาสถานภาพของแรงงาน (ตัวแปรตาม =1 หากมีงานทำ =0 หากว่างงาน) โดยพิจารณาจากอายุของแรงงาน อัตราการเจริญเติบโตทางเศรษฐกิจ ระดับความเชี่ยวชาญในฝีมือของแรงงาน

วงการการตลาด : ต้องการพิจารณาการตัดสินใจเลือกซื้อผลิตภัณฑ์ (ตัวแปรตาม =1 หากมีการซื้อ =0 หากไม่ซื้อ) โดยพิจารณาจากเพศ อายุ อาชีพของผู้บริโภค

วงการธุรกิจ : นักวิจัยใช้ Probit regression ในการศึกษาองค์กรธุรกิจที่ล้มละลาย โดยใช้ตัวอย่างของธุรกิจที่ล้มละลายจำนวน 96 รายและธุรกิจที่ไม่มีปัญหาจำนวน 3,880 ราย ตัวแปรอิสระที่ใช้คือ อัตราส่วนรายได้สุทธิต่อสินทรัพย์รวม อัตราส่วนของหนี้สินรวมต่อสินทรัพย์รวม และอัตราส่วนของสินทรัพย์หมุนเวียนต่อหนี้สินหมุนเวียน

วงการผลิตนมเพื่อดื่ม : นักวิจัยใช้ Probit regression ในการศึกษาพฤติกรรมการดื่มนมบรรจุกล่องกับนมที่ไม่มีการบรรจุกล่องของประชากรเมืองหนึ่งในประเทศตุรกี ตัวแปรอิสระได้แก่ ขนาดของครัวเรือน รายได้ของครอบครัว เหตุผลที่เลือกดื่มนม และราคา

วงการศึกษาระดับบัณฑิตศึกษา : นักการศึกษาจากมหาวิทยาลัยแห่งหนึ่งสนใจที่จะใช้ probit regression ในการศึกษาว่า คะแนน GRE (Graduate Record Exam score) เกรดเฉลี่ยของนักศึกษา (G.P.A.) และความมีชื่อเสียงของสถาบันการศึกษาในระดับปริญญาตรีที่นักศึกษาเรียน (Ivy League หรือ Top 10 ranking) มีผลต่อการรับนักศึกษาเข้าศึกษาต่อในระดับบัณฑิตศึกษา ส่วนตัวแปรตาม ได้แก่ ปฏิเสธ หรือ ตอบรับ

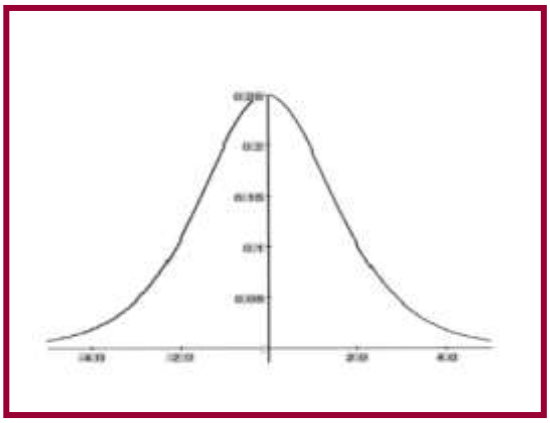
วงการเศรษฐศาสตร์แรงงาน : นักเศรษฐศาสตร์ด้านแรงงานสามารถใช้ probit regression เพื่อช่วยให้สามารถบอกได้ว่า ผลจากการมีการศึกษาในระดับที่สูงขึ้นจะมีส่วนช่วยเพิ่มความน่าจะเป็นที่จะถูกเลือกจ้าง ยิ่งไปกว่านั้นการนำตัวแปรอิสระอื่น ๆ เช่น อายุ ประสบการณ์การทำงาน และที่ตั้งทางภูมิศาสตร์ก็จะช่วยให้สามารถสร้างแบบจำลองที่สมบูรณ์และครอบคลุม

อย่างไรก็ตามแม้ logistic regression และ probit regression จะให้ผลใกล้เคียงกัน แต่วิธีการทางสถิติทั้งสองมีข้อสมมติฐานแตกต่างกันดังนี้

1.**ลักษณะการกระจายของค่าตัวแปรตาม**: ในขณะที่ logistic ใช้ logistic distribution probit regression จะสมมติว่าตัวแปร Y* ซึ่งเป็นตัวแปรแฝง(latent variable) ของตัวแปรตาม Y และเป็นlinear combination ของค่าตัวแปรอิสระในรูปแบบ $Y^*_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots\ldots \beta_k X_{ki} + u_i$ โดย Y* จะเป็นตัวแปรที่ไม่สามารถสังเกตเห็นได้ และมีความสัมพันธ์กับ Y คือ $Y_i=1$ ถ้า $Y_i^*>0$

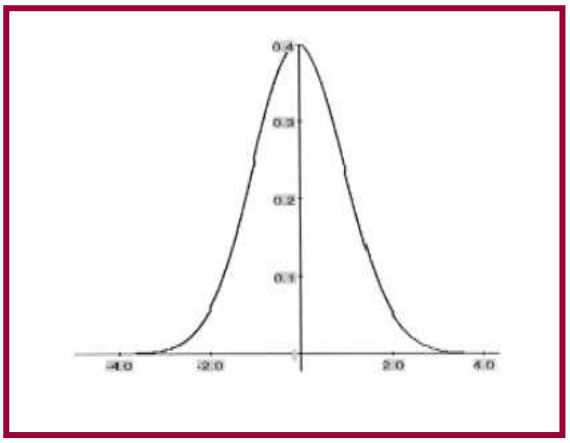
$=0$ ถ้า Y_i^*มีค่าเป็นอย่างอื่น

โดยที่ X_1 , X_2 ,……X_k เป็นตัวแปรอิสระ β's เป็น regression coefficients และ u เป็น random disturbance term ที่มีการกระจายแบบ Standard Normal distribution



probability density function ของ

logistic distribution

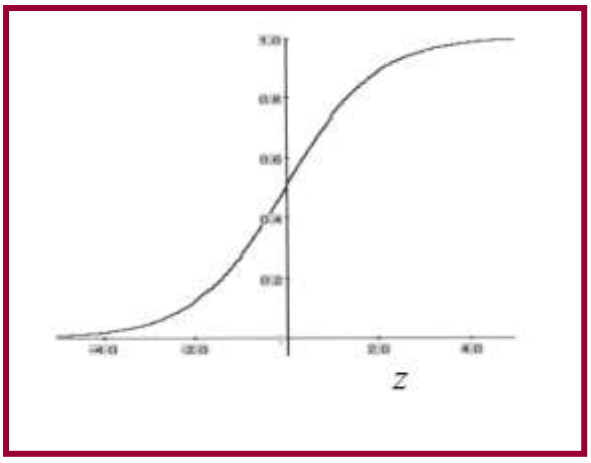


probability density function ของ

standard normal distribution

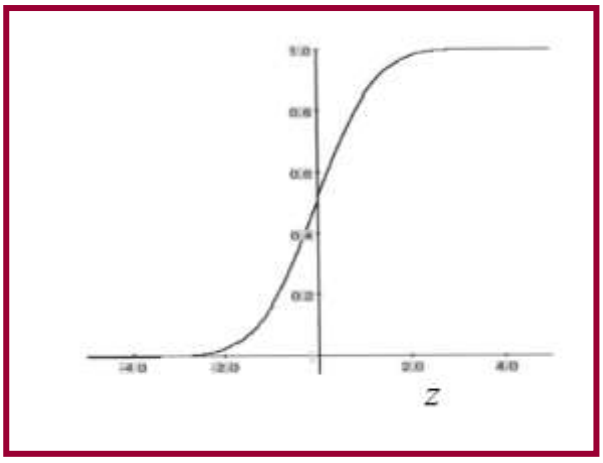
จากสมมติฐานนี้เอง หากมี case หลาย case ในกลุ่มตัวอย่างกระจายอยู่ในช่วงปลายข้างใดข้างหนึ่งมาก การกระจายของตัวแปรตามจะไม่มีลักษณะเป็น Normal และนักสถิติไม่ควรใช้ probit regression

2.**การ transform ตัวแปร** : ใน logistic ใช้ natural log ของodds ratio ในขณะที่การ transform ตัวแปรใน probit ใช้ inverse ของ Standard normal cumulative distribution function ผลจากการ transform ทำให้ได้เส้นที่มีลักษณะเป็นเส้นโค้งรูป s (S-shaped curve) โดยเส้น probit จะมีความชันมากกว่า เส้น logit อย่างไรก็ตาม ความแตกต่างระหว่างเส้นสองเส้นมีน้อย



Cumulative distribution function ของ

logistic regression



Cumulative distribution functionของ

standard normal distribution

3.**Coefficients:** regression coefficients ของ logit จะแตกต่างจาก coefficients ของ probit โดย coefficients ของ logit จะเป็นประมาณ 1.6-1.8 เท่าของ coefficients ของ probit

Coefficients ของ logistic สามารถตีความได้ง่ายกว่า probit regression โดย coefficients ของ logits จะบอกการเปลี่ยนแปลงใน log-odds ของตัวแปรตาม ในขณะที่ coefficients ของ probit จะบอกผลของตัวแปรอิสระที่มีต่อ Z-score ของตัวแปรตาม ยิ่งไปกว่านั้นความน่าจะเป็นของตัวแปรตามไม่ได้มีความสัมพันธ์เชิงเส้นตรงกับ Z-score แต่ขึ้นอยู่กับค่าตัวแปรอิสระด้วย ด้วยเหตุผลนี้เอง logit จึงมีผู้นิยมใช้มากกว่า probit แต่ก็ขึ้นอยู่กับความชอบ ตลอดจนศาสตร์ในสาขาที่นักสถิติสังกัดด้วย อย่างไรก็ตาม ผลที่ได้จาก logit และ probit จะให้ข้อสรุปทางสถิติคล้ายคลึงกัน

ข้อควรระวังในการตีความ: ในการตีความผลที่ได้จาก (binary) logistic regression หรือ (binary) probit regression จะพิจารณาเครื่องหมายหน้า regression coefficients มากกว่าการพิจารณาขนาดของ regression coefficients ทั้งนี้เนื่องจาก scale ของ regression coefficients ที่ใช้ใน logistic และ probit จะต่างกัน โดยหากเครื่องหมายหน้า regression coefficient เป็นบวก มีความหมายว่า การเพิ่มขึ้นของตัวแปรอิสระมีผลทำให้โอกาสที่ค่าของตัวแปรตาม(ที่กำหนดค่า =1) จะเกิดเพิ่มสูงขึ้น และการลดลงของตัวแปรอิสระมีผลทำให้โอกาสที่ค่าของตัวแปรตาม(ที่กำหนดค่า =1)จะเกิดลดน้อยลง

การใช้โปรแกรมสำเร็จรูปด้วย probit regression

สามารถทำได้ 3 วิธีดังนี้

1. ใช้ GENLIN syntax command โดยมี format ดังนี้

```
genlin dependent variable (reference=0) with ind.v1 ind.v2 ind.v3
/model ind.v1 ind.v2 ind.v3
distribution=binomial
link=probit
/print cps history fit solution.
```

<u>คำอธิบาย</u>

genlin คำสั่งเรียก generalized linear model command

dependent variable: ชื่อตัวแปรตาม

indv: ชื่อตัวแปรอิสระ

model: ตามด้วยชื่อตัวแปรอิสระที่ใช้ในแบบจำลอง

distribution = ตามด้วยรูปแบบการกระจายของผลที่เกิดขึ้นในแต่ละครั้ง ซึ่งในที่นี้เป็น binomial distribution

link= ฟังก์ชั่นที่ลิงค์ ในที่นี้ใช้ probit function

print: คำสั่งให้แสดง output

cps: case processing summary

history: แสดงผลที่เกิดขึ้นตามลำดับ

fit: goodness of fit

solution: แสดงผลที่ได้จากการ run โปรแกรมสำเร็จรูป

2. ใช้ PLUM syntax command โดยมี format ดังนี้

```
plum dependent variable with indv.1 indv.2 indv.3
/link = probit
/print = parameter summary
```

คำอธิบาย

plum คำสั่งเรียก polytomous universal model command

parameter: parameter estimates

summary: สรุปผลทางสถิติอื่น ๆ

โปรดสังเกต : PLUM syntax command เป็น syntax command ที่ใช้ในการวิเคราะห์ ordinal regression โดยเฉพาะ แต่แบบจำลองที่มีตัวแปรตามที่แบ่งออกได้เป็นสองค่า เช่น binary probit regression ถือเป็นกรณีพิเศษของ ordinal regression เพียงแต่เราต้องใช้ link function ให้ถูกต้อง (ในที่นี้ ใช้ probit function)

3. ใช้ GZLM-drop down menu ซึ่งเป็น menu หนึ่งใน generalized linear models

Computer output ที่ได้จากโปรแกรมสำเร็จรูป: ประกอบด้วย

1. Model information

แสดงให้เห็นชื่อของตัวแปรตาม Probability function ที่ใช้และ Link function

2. Case processing summary

แสดงให้เห็นขนาดของกลุ่มตัวอย่าง

3. Categorical Variable Information

แสดงให้เห็นการกระจายของกลุ่มตัวอย่างใน factor space

4. Continuous variable information

แสดงให้เห็นค่าสถิติเชิงพรรณนาของตัวแปรอิสระ

5. Goodness of fit table

ให้ข้อมูลเกี่ยวกับ Deviance Chi-square , Pearson Chi-square, AIC ,BIC

6.Omnibus test

แสดงการเปรียบเทียบผลที่เกิดจากการใช้ full model กับผลที่เกิดจาก null model ว่าสามารถลดความผันผวนของตัวแปรตามที่ไม่สามารถอธิบายได้ลงในระดับที่มีนัยสำคัญทางสถิติหรือไม่ โดยใช้สถิติ Chi-square

Chi-Square หรือ log likelihood Chi-Square test เป็นการทดสอบว่า อย่างน้อยมีregression coefficient ของตัวแปรอิสระใน model บางตัวที่มีค่าไม่ใช่ ศูนย์ (0) LR Chi-square statistic คำนวณจากผลต่างของ -2LL ของ null model และ fitted model

Df=degree of freedom แสดงจำนวนตัวแปรอิสระที่ใช้ในการพยากรณ์

7. Model fitting Table

แสดงการเปรียบเทียบ -2 Log-likelihood ของ null model(intercept only model) และ full model (fitted model)ตลอดจนผลต่างว่ามีนัยสำคัญหรือไม่ ผลต่างของ-2Log Likelihoodระหว่าง model แสดงให้เห็นผลจากการพยากรณ์ที่มีความแม่นยำเพิ่มมากขึ้นจากการนำตัวแปรอิสระเข้ามาช่วยพยากรณ์ ยิ่งผลต่างของ-2LL ระหว่างสองmodelยิ่งสูง ก็เท่ากับว่าการนำตัวแปรอิสระเข้ามาช่วยในการพยากรณ์ประสบความสำเร็จมากขึ้นเท่านั้น

8.Tests of model effects

แสดงให้เห็นว่าตัวแปรอิสระตัวใดมีส่วนสำคัญในการอธิบายความผันผวนของตัวแปรตาม

Sig. ใช้พิจารณาในการทดสอบสมมติฐานว่า ค่าของregression coefficients ของตัวแปรอิสระทุกตัวมีค่าเป็นศูนย์หรือไม่ หากมีค่ามากกว่า ค่าα ที่กำหนดให้ จึงจะยอมรับสมมติฐานว่า ค่าของ regression coefficients ของตัวแปรอิสระทุกตัวมีค่าเป็นศูนย์ หรืออีกนัยหนึ่ง ตัวแปรอิสระไม่ได้มีส่วนในการช่วยพยากรณ์ค่าของตัวแปรตามเลย

9. Pseudo-R-square

แสดงให้เห็นค่าของ R-square เทียม ซึ่งให้ภาพคร่าว ๆเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม โดยPseudo R-square เป็นความพยายามของนักสถิติที่ต้องการค่าสถิติในลักษณะเดียวกับ R-square ที่ใช้ในการวิเคราะห์การถดถอยพหุคูณ อย่างไรก็ตาม การตีความค่า R-square ที่ได้นี้ (มีถึง 3 ค่า ได้แก่ R-square ของ Cox& Snell ของ Nagelkerke และของ McFadden) จะต้องกระทำด้วยความระมัดระวัง เนื่องจากไม่ละเอียดเท่ากับค่า R-square ที่ใช้ในการถดถอยพหุคูณ

10.Parameter Estimates

แสดงให้เห็นค่า regression coefficients หน้าตัวแปรอิสระ อย่างไรก็ตามเนื่องจากการตีความของ probit regression ไม่ได้มีความหมายตรงไปตรงมาเหมือนกับการตีความใน logistic regression หรือ ordinary linear regression ทำให้เราไม่สามารถระบุตัวเลขความน่าจะเป็นที่เพิ่มขึ้นได้แน่นอน

Estimate: แสดงค่าของ regression coefficients ของตัวแปรอิสระที่ใช้ในการกำหนดค่าของ probability โดยฟังก์ชั่นของการคำนวณ probability จะใช้ Cumulative Distribution function ของ Standard normal distribution

อนึ่งใคร่ขอตั้งข้อสังเกตุว่า การตีความregression coefficients จะไม่ตรงไปตรงมาเหมือนกับการตีความ regression coefficients ใน linear regression หรือ logistic regression ผลจากการเปลี่ยนแปลงใน probability จากการที่ค่าของตัวแปรอิสระหนึ่งๆเปลี่ยนแปลงขึ้นอยู่กับค่าของตัวแปรอิสระอื่น ๆ และขึ้นอยู่กับค่าเริ่มต้นของตัวแปรที่เรากำลังพิจารณาอยู่

ใน probit regression หาก regression coefficients เป็นบวก นั่นหมายความว่าการเพิ่มค่าของตัวแปรอิสระนั้นมีผลทำให้ค่าของ probability สูงขึ้น และหาก regression coefficient เป็นลบ นั่นหมายความว่าการเพิ่มค่าของตัวแปรอิสระนั้นมีผลทำให้ค่าของ probability ลดลง

Regression coefficients ใน probit regression จะบอกการเปลี่ยนแปลงในค่าของ Z-score (บางทีเรียกว่า probit index) เมื่อค่าของตัวแปรอิสะเปลี่ยนแปลงไปหนึ่งหน่วย

Wald statistic: หาได้โดยเอาค่าของ regression coefficient(ค่า Estimate) หารด้วย standard error แล้วยกกำลังสอง Wald statistic จะมีการกระจายแบบChi-square

Wald statistic จะใช้บอกว่าตัวแปรแต่ละตัวมีค่าแตกต่างจากศูนย์อย่างมีนัยสำคัญหรือไม่ เมื่อมีตัวแปรอิสระอื่น ๆอยู่ใน model

Contribution this issue: ดร. ดนัย ปัตตพงศ์

* 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)