16

Logistic regression

เป็นเทคนิคทางสถิติที่ใช้ตัวแปรอิสระในการพยากรณ์ค่าของตัวแปรตามหรือตัวแปรผล (outcome variable) โดยตัวแปรตามอาจมีได้สองค่าหรือหลายค่า ในกรณีที่มีเพียงสองค่า เทคนิคที่ใช้พยากรณ์เรียกว่า (binary) logistic regression หากมีมากกว่าสองค่า เทคนิคที่ใช้พยากรณ์เรียกว่า multinomial logistic regression ในกรณีที่ตัวแปรตามเป็นตัวแปรอันดับ (ordinal variable) เทคนิคที่ใช้พยากรณ์เรียกว่า Ordered logistic regression ในเอกสารนี้เราจะมุ่งอธิบายเฉพาะ (binary) logistic regression หรือที่เรียกกันสั้นๆว่า logistic regression

เพื่อให้เห็นลักษณะของปัญหาที่มีการนำเอา logistic regression (LR) ไปใช้ประโยชน์ในหลากหลายวงการ เราลองมาพิจารณาตัวอย่างดังต่อไปนี้

วงการสาธารณสุข

นักวิจัยและบุคลากรทางการแพทย์ สามารถนำ LR ไปใช้พยากรณ์ว่าบุคคลหนึ่งๆมีโอกาสที่จะติดเชื้อได้หรือไม่ โดนอาศัยข้อมูลจากจำนวนผู้ที่ติดเชื้อ อายุของบุคคลนั้น ๆ สุขภาพ สภาพแวดล้อม ความสามารถในการเข้าถึงทรัพยากรทางการแพทย์

วงการการตลาด

ธุรกิจสามารถใช้ LR ในการพยากรณ์ว่าลูกค้าจะตีจากเราไปโดยอาศัยประวัติการติดต่อ ข้อมูลประชากรศาสตร์ และลักษณะของการใช้สินค้าและบริการ

วงการโทรคมนาคม

กิจการโทรคมนาคมสามารถพยากรณ์โอกาสที่ลูกค้าจะ upgrade บริการที่ให้หรือเปลี่ยนตัวผู้ให้บริการโดยพิจารณาจากลักษณะการใช้บริการตลอดจนปฏิสัมพันธ์ที่มีในการให้บริการแก่ลูกค้า

สถาบันการเงินกับการจัดอันดับลูกหนี้

สถาบันการเงินใช้ LR ในการประเมินโอกาสที่ลูกค้าจะผิดนัดชำระหนี้โดยอาศัยประวัติการมีสินเชื่อและดัชนีด้านการเงินอื่น ๆ

วงการบัตรเครดิต

ธุรกิจที่ออกบัตรเครดิตสามารถใช้ LR ในการกำหนดธุรกรรมการใช้บัตรเครดิตที่ส่อว่าจะทุจริต โดยพิจารณาจากจำนวนเงินที่เกี่ยวข้องในการทำธุรกรรม ระดับของcredit scoreของผู้ถือบัตร

วงการธุรกิจ

ใช้ LR ในการกำหนดว่า email ที่เข้ามาจะเป็น spam หรือไม่โดยพิจารณาจากจำนวนคำและแหล่งประเทศต้นทาง

วงการการศึกษา

นักการศึกษาสามารถใช้ LR ในการพยากรณ์โอกาสที่นักศึกษาจะได้รับการคัดเลือกให้เข้าเรียนในมหาวิทยาลัยโดยพิจารณาจากคะแนน GPA , ACT และจำนวนการเรียนวิชาเสริม

วงการด้านการแพทย์

นักวิจัยด้านการแพทย์สามารถใช้ LR ในแบบจำลองที่จะบอกได้ว่าการเปลี่ยนแปลงในการออกกำลังกายและน้ำหนักจะมีผลต่อโอกาสที่บุคคลผู้หนึ่งจะมีอาการหัวใจวาย

วงการโลจิสติกส์

ใช้ LR ในการพยากรณ์เวลา ตลอดจนอุปกรณ์ที่จำเป็นต้องมีการบำรุงรักษาโดยพิจารณาจากประวัติการบำรุงรักษาและข้อมูลอื่น ๆที่เกี่ยวข้อง เพื่อดำเนินการให้มีแผนบำรุงรักษาล่วงหน้าที่มีประสิทธิภาพ

ธุรกิจกับการพยากรณ์อุปสงค์

LR สามารถช่วยในการพยากรณ์ลักษณะอุปสงค์โดยพิจารณาจากประวัติข้อมูล ทำให้องค์กรธุรกิจสามารถลดต้นทุนการสต๊อกสินค้าและช่วยให้สามารถส่งมอบสินค้าได้ทันเวลา

วงการเกม

ใช้ LR ในการเสนอแนะผู้เล่นในระหว่างเล่นเกมว่าควรจะมีการเติมเงินในเกม เพื่อเพิ่มความสามารถพิเศษให้กับตัวละครในเกม การกำหนดภาพลักษณ์ เครื่องมือที่จำเป็นในการต่อกรกับคู่แข่ง ตลอดจนสื่อสารกับผู้เล่นคนอื่น ๆ โดยอาศัยข้อมูลของผู้เล่นเกมเองกับข้อมูลของผู้เล่นเกมคนอื่นๆซึ่งรวมถึงสิ่งที่ชื่นชอบ

วงการวิศวกรรม

LR สามารถใช้ในการพยากรณ์ว่าเครื่องจักรหรือชิ้นส่วนจะล้มเหลวโดยอาศัยข้อมูลเกี่ยวกับอุณหภูมิ ระดับการสั่น และจำนวนชั่วโมงของการทำงาน

วงการแพทย์

LR ใช้ในการวินิจฉัยโรคเบาหวาน โรคหัวใจโดยพิจารณาปัจจัยเสี่ยงเช่น อายุ น้ำหนัก ประวัติครอบครัว หรือนำ LR ไปใช้ในการประเมินประสิทธิภาพการเยียวยารักษาโดยอาศัยตัวแปรที่เกี่ยวข้องกับการบำบัดรักษาต่าง ๆ

วงการสถาบันการเงิน

ใช้ LR ในการพยากรณ์ธุรกรรมที่มีแววส่อว่าจะทุจริตโดยพิจารณาจากจำนวนเงินที่เกี่ยวข้อง สถานที่ทำธุรกรรม และเวลาของการทำธุรกรรม

วงการธุรกิจ

ธุรกิจต่าง ๆโดยเฉพาะอย่างยิ่ง ธุรกิจโทรคมนาคม และบริการรับสมัครเป็นสมาชิก ใช้ LR ในการทำนายลูกค้าที่หันไปใช้สินค้าหรือบริการอื่น ๆ ตัวอย่างเช่น Booking.com ใช้ LR ในการพยากรณ์พฤติกรรมลูกค้า เช่นการเปลี่ยนวันที่เดินทาง โดยไม่จำเป็นต้องมีข้อมูลของลูกค้าเป็นการเฉพาะ

วงการสาธารณสุข

นักการสาธารณสุขใช้ LR ในการพยากรณ์การเกิดโรค โดยเฉพาะอย่างยิ่งในการเปิดเผยความเชื่อมโยงระหว่าง microRNA กับยีนส์ ช่วยเพิ่มความแม่นยำในผลตรวจเลือด ทำให้ตรวจเจอโรคที่เกี่ยวกับยีนส์ได้แต่เนิ่น ๆ

เพื่อให้เห็นว่า logistic regression ซึ่งใช้ในการพยากรณ์ค่าของตัวแปรตาม (สมมติว่า Y) ในรูปของความน่าจะเป็นที่มีค่าอยู่ระหว่าง 0 (ศูนย์) และ 1 มีแนวคิดทางสถิติอย่างไร ลองพิจารณาตัวอย่างต่อไปนี้

ตัวอย่างที่หนึ่ง : ในการออกเสียงโหวตเลือกพรรคการเมือง เราอาจกำหนดให้

$$Y=\begin{cases}1 \text{ หากผู้โหวตเลือกพรรคสุขสันต์}\\ 0 \text{ หากผู้โหวตเลือกพรรคการเมืองอื่น ๆ}\end{cases}$$

ส่วนตัวแปรอิสระ (X1, X2 , X3,….) ที่จะใช้พยากรณ์ความน่าจะเป็นที่ผู้สมัครจะเลือกพรรคสุขสันต์อาจเป็นเพศ อายุ ระดับการศึกษา ภาคที่สังกัดอยู่ ฯลฯ

ตัวอย่างที่สอง : ในการวินิจฉัยระดับความดันโลหิตของผู้ป่วยที่มีอาการของโรคความดันโลหิตสูง เราอาจกำหนดให้

$$Y = \begin{cases} 1 \text{ หากเป็นโรคความดันโลหิตสูง} \\ 0 \text{ หากความดันโลหิตอยู่ในระดับปกติ} \end{cases}$$

ส่วนตัวแปรอิสระ (X1, X2 , X3,….)ที่จะใช้พยากรณ์ความน่าจะเป็นที่ผู้ป่วยจะเป็นโรคความดันโลหิตสูงอาจเป็นเพศ อายุ มีพฤติกรรมการสูบหรือไม่สูบบุหรี่ ฯลฯ

ตัวอย่างที่สาม : ในการประเมินความสำเร็จของการบำบัดรักษาอย่างหนึ่ง เราอาจกำหนดให้

$$Y = \begin{cases} 1 \text{ หากการบำบัดรักษาประสบผลสำเร็จ} \\ 0 \text{ หากการบำบัดรักษาไม่ประสบผลสำเร็จ} \end{cases}$$

ส่วนตัวแปรอิสระ (X1 , X2 , X3,…..) ที่จะใช้พยากรณ์ความน่าจะเป็นที่การรักษาจะประสบผลสำเร็จอาจเป็นขนาดของยา เพศ อายุ น้ำหนัก ความหนักเบาของอาการที่เป็น ฯลฯ

เนื่องจากตัวแปรตามมีได้สองค่าเท่านั้น ดังนั้นเราไม่สามารถใช้เทคนิคการวิเคราะห์การถดถอย (regression analysis) มาใช้ได้ เนื่องจากตัวแปรตามไม่ได้มีการกระจายแบบ Normal(สมมติฐานสำคัญของการวิเคราะห์การถดถอย) ยิ่งไปกว่านั้นความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระไม่ได้มีความสัมพันธ์เชิงเส้นตรง (อีกสมมติฐานที่สำคัญของการวิเคราะห์การถดถอย) สิ่งที่สามารถอธิบายให้เห็นได้ว่า เราไม่สามารถใช้การวิเคราะห์การถดถอย มาใช้ได้ สามารถพิจารณาจากสมการดังต่อไปนี้

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3$$

ในที่นี้ถ้ากำหนดว่า Y เป็นตัวแปรตามซึ่งมีค่าเพียงสองค่า ส่วน X_1 , X_2 , X_3 เป็นตัวแปรอิสระ ส่วน a_0 , a_1, a_2, a_3 เป็นค่าคงที่ จะเห็นได้ว่าเนื่องจากฝั่งขวาของสมการอาจมีค่าใดก็ได้ระหว่าง $-\infty$ ถึง $+\infty$ ดังนั้นเราจึงไม่มีทางกำหนด Y ให้มีค่าเป็นสองค่าได้เลย นอกจากจะต้องมีค่าอยู่ระหว่างลบอินฟินิตี้ ($-\infty$) ถึงบวกอินฟินิตี้ ($+\infty$) เช่นกัน นั่นก็คือเทคนิคการวิเคราะห์การถดถอยไม่สามารถนำมาใช้ได้ในสถานการณ์เช่นนี้

นักสถิติได้คิดค้นหาทางออก โดยแทนที่จะพยากรณ์ค่าของ Y กลับกำหนดในรูปความน่าจะเป็นที่ Y จะมีค่าใดค่าหนึ่งแทน ตัวอย่างเช่นถ้ากำหนดว่าเพศของเด็กที่เกิดมาเป็นชาย ก็จะให้รหัส 1 แต่ถ้าเกิดมาเป็นหญิงก็จะให้รหัส 0 ดังนี้เป็นต้น นักสถิติจะหันมาให้ความสำคัญกับการใช้ตัวแปรอิสระ(อายุบิดา อายุมารดา น้ำหนักบิดา น้ำหนักมารดา ฯลฯ) ในการพยากรณ์ความน่าจะเป็นที่เพศของเด็กที่เกิดมาจะเป็นชาย

เนื่องจากค่าของความน่าจะเป็นมีค่าอยู่ระหว่าง 0 และ 1 ในขณะที่ค่าของตัวแปรอิสระอาจมีค่าใดก็ได้ที่อยู่ระหว่าง $-\infty$ และ $+\infty$ ดังนั้นการจะผูกโยงเข้ากันได้ จะต้องกระทำในลักษณะของ Log –Odd แสดงได้ดังนี้

$$\text{Log}_e(\text{odd- ratio}) = b_0 + b_1X_1 + b_2X_2 + \ldots b_nX_n \qquad (1)$$

โดยที่ odd- ratio = Probability(success)/Probability(failure)

และถ้ากำหนดให้ π=Probability(success) (1) สามารถเขียนใหม่ได้เป็น

$$\frac{\pi}{1-\pi} = e^{(b_O + b_1 X1 + b_2 X2 + \ldots.. b_n Xn)} \quad (2)$$

หรืออยู่ในรูปต่อไปนี้

$$\pi = \frac{1}{1 + e^{-(b_o + b_1 X1 + b_2 X2 + \ldots.. b_n Xn)}} \quad (3)$$

ความหมายก็คือ (1)=(2)=(3)

แต่ (1) แสดงให้เห็นตัวแปรตาม (log-odd) ที่มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ

สมมติฐานที่จำเป็นในการใช้เทคนิคนี้ :-

1. ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระอาจเป็น linear หรือไม่เป็นก็ได้ แต่ log-odd ของตัวแปรตามและตัวแปรอิสระต้องเป็น linear
2. ไม่ได้ตั้งสมมติฐานว่า error term มีการกระจายแบบ Normal
3. ตัวแปรตามต้องเป็นตัวแปรที่มีค่าได้สองค่า
4. Error terms ต้องเป็นอิสระต่อกัน ไม่มีความสัมพันธ์กัน หรือเกิดจากการทดลองซ้ำหรือกลุ่มคนที่มีถิ่นฐานทางภูมิศาสตร์ใกล้กัน
5. ต้องไม่มีข้อมูลสุดโต่งที่หลุดจากกลุ่ม (outlier)
6. เนื่องจากการประมาณค่าของ logistic regression coefficient ใช้วิธีการของ Maximum likelihood Estimation ดังนั้นขนาดของกลุ่มตัวอย่างต้องมีมากพอ
7. จะต้องไม่มี multicollinearity ในระหว่างตัวแปรอิสระ

ในส่วนของตัวแปรอิสระหรือตัวแปรที่ใช้ในการพยากรณ์ (predictor variables) นั้น อาจอยู่ในรูปตัวแปรที่มีค่าต่อเนื่อง (continuous variable) หรือในรูปตัวแปรที่แสดงประเภท (categorical variable) หรือ ทั้งสองรูป

กลไกการวิเคราะห์แบบlogistic regressionจะมีลักษณะเป็นขั้นตอน (stepwise) โดยจะเริ่มจากการพิจารณาผลทางสถิติโดยไม่นำเอาตัวแปรอิสระใด ๆ เข้ามาใช้ในการพยากรณ์ก่อน จากนั้นจึงนำตัวแปรอิสระเข้ามา อาจจะเข้ามาพร้อมกันหมดในคราวเดียวกัน หรือเข้ามาทีละตัวตามเงื่อนไขทางสถิติที่ผู้วิเคราะห์เป็นผู้กำหนดให้ ในแต่ละขั้นตอนจะมีการคำนวณดัชนีชี้วัดว่า ยังมีความคลาดเคลื่อนที่ยังไม่สามารถอธิบายได้อีกหรือไม่ ดัชนีชี้วัดที่ใช้นี้เรียกว่า log-likelihood แสดงได้โดยสูตรต่อไปนี้

$$\text{Log-likelihood} = \sum[y_i \ln(p(Y_i)) + (1 - y_i)\ln[1-p(y_i)]] \quad (4)$$

โดยที่ $y_i = 1$ ถ้า success หรือ = 0 ถ้า failure และเป็นค่าที่ได้จากการสังเกต

$P(y_i)$ = ความน่าจะเป็นที่ Y_i จะเกิดซึ่งเป็นค่าที่เป็นผลจากการพยากรณ์ของแบบจำลอง

ในขั้นตอนแรกที่ยังไม่มีการนำเอาตัวแปรอิสระใดมาช่วยการพยากรณ์ ค่าของ log-likelihood จะมีค่าสูง แสดงให้เห็นว่าหากยังไม่มีการนำตัวแปรอิสระตัวใดเข้ามา ความสามารถในการพยากรณ์จะต่ำและมีความคลาดเคลื่อนที่ยังไม่สามารถอธิบายได้อยู่ในระดับสูง ต่อเมื่อนำตัวแปรอิสระเข้ามา log-likelihood จะมีค่าเล็กลง เป็นนัยให้ทราบว่า ความคลาดเคลื่อนที่ยังไม่สามารถอธิบายได้มีขนาดเล็กลง

ผลต่างในความคลาดเคลื่อนระหว่างกรณีที่ยังไม่มีการนำตัวแปรใดเข้ามา(กรณีที่ใช้เป็นเกณฑ์ฐาน) กับกรณีที่มีการนำตัวแปรตัวใดตัวหนึ่งหรือทั้งหมดเข้ามาอาจมีค่าต่ำบ่งบอกให้เห็นว่า การนำตัวแปรอิสระเข้ามา มีส่วนช่วยในการพยากรณ์ได้บ้าง แต่หากมีค่าสูง ย่อมหมายความว่า การนำตัวแปรอิสระเข้ามามีส่วนช่วยในการพยากรณ์ได้มาก นักสถิติมีวิธีวัดสัดส่วนความแตกต่างในความคลาดเคลื่อนที่เปลี่ยนไปจากเดิมนี้ว่ามีนัยสำคัญทางสถิติหรือไม่โดยใช้ค่าสถิติ χ^2 ที่นักสถิติคุ้นเคยกัน คำนวณได้โดยใช้สูตรดังต่อไปนี้

$$\chi^2 = 2\,[LL(New)-LL(Baseline)] \qquad (5)$$

$$\text{degree of freedom} = k_{new} - k_{baseline}$$

โดยที่ χ^2 คือค่าสถิติไคว์สแควร์

LL(New) = log-likelihood คำนวณจากการนำตัวแปรอิสระเข้ามา

LL(Baseline) = log-likelihood คำนวณจากการไม่มีตัวแปรอิสระใด ๆเข้ามา

$k_{baseline}$ = degree of freedom จากการไม่มีตัวแปรอิสระใด ๆเข้ามาซึ่งจะมีค่าเท่ากับ 1 เนื่องจากต้องมีการคำนวณค่าคงที่ (b_O) ตาม (1)

k_{new} = degree of freedom จากการนำตัวแปรอิสระเข้ามา ซึ่งจะมีค่าเท่ากับ 1 บวกกับจำนวนตัวแปรอิสระที่นำเข้ามา

ดัชนีชี้วัดความมีประสิทธิภาพในการพยากรณ์ของ logistic regression

มีหลายดัชนี ซึ่งต้องใช้ประกอบกัน แต่ไม่จำเป็นต้องใช้ทั้งหมด ที่เป็นเช่นนี้ก็เพราะการตีความผลที่ได้จากlogistic regression อาจไม่สามารถกำหนดตีความผลออกมาได้ชัดเจนกระจ่างแจ้งเหมือนกับผลที่ได้จากการวิเคราะห์การถดถอย (regression analysis) กล่าวอีกนัยหนึ่งก็คือ การตีความผลที่ได้จาก logistic regression จะต้องกระทำด้วยความระมัดระวัง ไม่สามารถสรุปได้ด้วยค่าของดัชนีใดดัชนีหนึ่ง แต่หากหลายดัชนีชี้ไปในทิศทางเดียวกัน เมื่อนั้น ผู้วิจัยจะมีความมั่นใจในการสรุปผลและตีความผลที่ได้จาก logistic regression

Classification table: ได้แก่พิจารณาผลจากสัดส่วนของตัวอย่างที่ได้จากการสังเกตซึ่งมีการพยากรณ์ได้อย่างถูกต้องซึ่งมีแสดงผลในรูปตาราง ยิ่งสัดส่วนที่มีความถูกต้องในการพยากรณ์ยิ่งเข้าใกล้ 100 เท่าใด ก็ยิ่งแสดงให้เห็นความสำเร็จในการใช้ logistic regression ในการพยากรณ์มากขึ้นเท่านั้น ลองพิจารณาผลจากการใช้ logistic regression สำหรับกรณี 2 กรณีดังต่อไปนี้

	ผลจากการพยากรณ์		
ที่เกิดขึ้นจริง	success	failure	ถูกต้องร้อยละ
success	20	80	20
failure	80	20	20
รวมร้อยละ			20

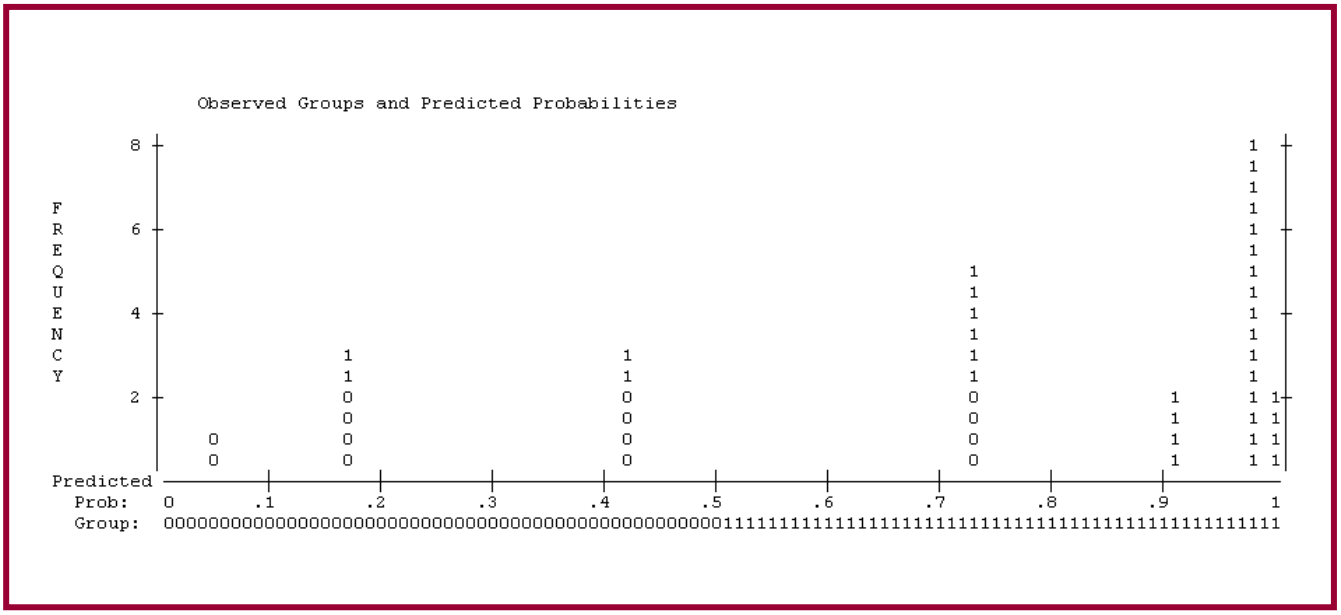
กรณีที่ 1

	ผลจากการพยากรณ์		
ที่เกิดขึ้นจริง	success	failure	ถูกต้องร้อยละ
success	80	20	80
failure	20	80	80
รวมร้อยละ			80

กรณีที่ 2

จะเห็นได้ว่า logistic regression มีความถูกต้องของการพยากรณ์ในกรณีที่ 2 สูงกว่าในกรณีที่ 1

Classification plot : ถ้ากำหนดว่า 1= success และ 0= failure เราจะพิจารณาจาก classification plot ว่ามีกรณีที่ success ถูก misclassify ให้ไปอยู่กลุ่ม failure (กลุ่มที่ predicted probability มีค่าน้อยกว่า 0.5) มากน้อยเท่าใด ถ้ามีน้อยแสดงว่า ความถูกต้องในการพยากรณ์มีสูง ขณะเดียวกันก็ต้องพิจารณาว่า failure ถูก misclassify ให้ไปอยู่ในกลุ่ม success (กลุ่มที่ predicted probability มีค่ามากกว่า 0.5) มากน้อยเท่าใด ลองพิจารณาตัวอย่างในภาพด้านล่างประกอบ



จากภาพ กลุ่ม success (กลุ่มที่มีรหัสเท่ากับ 1) ถูก misclassify ให้ไปอยู่ในกลุ่ม failure (กลุ่มที่มีรหัสเท่ากับ 0) มีจำนวนเท่ากับ 4 ส่วนกลุ่ม failure ที่ถูก misclassify ให้ไปอยู่ในกลุ่ม success มีจำนวนเท่ากับ 4 ยิ่งจำนวนที่ถูก misclassify มีน้อยเท่าใด ยิ่งแสดงความแม่นยำของการพยากรณ์โดยใช้ logistic regression เท่านั้น

Likelihood Ratio test : เป็นการเปรียบเทียบผลต่างระหว่างค่าของสถิติ -2LL ของกรณีที่ไม่มีตัวแปรอิสระ (baseline) กับกรณีที่มีตัวแปรอิสระเข้ามาช่วยในการพยากรณ์ ผลต่างบ่งบอกการเปลี่ยนแปลงใน -2LL อันเป็นผลจากการนำตัวแปรอิสระเข้ามา หากผลต่างมีค่าสูง แสดงให้เห็นว่า การเพิ่มตัวแปรอิสระเข้ามามีส่วนช่วยลดความคลาดเคลื่อนที่ยังไม่สามารถอธิบายได้โดย model ลงในสัดส่วนที่อาจมีนัยสำคัญทางสถิติ การที่จะกำหนดว่าผลต่างที่ลดลงนี้มีนัยสำคัญหรือไม่ ใช้การทดสอบสถิติโดยใช้สูตรที่ (5) หากพบว่ามีนัยสำคัญ หมายความว่า การเพิ่มตัวแปรอิสระเข้ามาเกิดประโยชน์ ผลการทดสอบโดยใช้ Likelihood ratio test จะมีปรากฏใน ตาราง Omnibus tests of model coefficients

Wald test: ในการวิเคราะห์การถดถอย (regression analysis) จะมีการทดสอบโดยใช้ t-statistic เพื่อดูว่า coefficient ของตัวแปรอิสระมีค่าต่างจากศูนย์ (หากเป็นศูนย์แสดงว่า ตัวแปรอิสระนั้นไม่ช่วยในการพยากรณ์ค่าของตัวแปรตาม) หรือไม่ สูตรที่ใช้กันก็คือ

$$\frac{b_i}{S_{bi}} \qquad (6)$$

โดยที่ b_i = coefficient หน้าตัวแปร X_i

S_{bi} = standard error of estimate

หากเรายกกำลังสองสูตรที่ (6) เราจะได้ Wald statistic ซึ่งมีการกระจายแบบ Chi-square ที่ degree of freedom =1 ข้อมูลเกี่ยวกับ Wald statistic จะมีแสดงในตาราง Variables in the equation หากพบว่าค่าสถิติมีนัยสำคัญ แสดงว่า coefficient หน้าตัวแปรอิสระที่นำเข้ามาช่วยในการพยากรณ์ใน logistic regression แตกต่างจากศูนย์อย่างมีนัยสำคัญ

R^2 : ในการวิเคราะห์การถดถอย(regression analysis) coefficient of (multiple) determination หรือ R^2 จะบ่งบอกสัดส่วนความผันผวนในตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระ ใน logistic regression ก็เช่นกัน จะมีการคำนวณค่าของ R^2 ซึ่งจะให้ข้อมูลเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเช่นเดียวกัน โดยถ้าค่าที่ได้มีค่าสูง ย่อมแสดงว่า การพยากรณ์โดยใช้ logistic regression ประสบผลสำเร็จ แต่ผู้วิเคราะห์ต้องระมัดระวังอย่าด่วนสรุปว่าตัวแปรอิสระสามารถอธิบายความแปรปรวนของตัวแปรตามได้เป็นสัดส่วนหนึ่งที่กำหนด ทั้งนี้เนื่องจาก R^2 ที่ใช้ใน logistic regression จะมีการคำนวณที่แตกต่างไปจากการคำนวณค่าของ R^2 ในการวิเคราะห์การถดถอย

R^2 ที่ใช้ใน logistic regression มีสูตรแตกต่างกันดังนี้

Hosmer & Lemeshow: $R^2 = \frac{-2LL(Model)}{-2LL(Baseline)}$

Cox & Snell : $R^2_{CS} = 1 - e^{[-2/n(LL(Model)-LL(Baseline)\}}$

เนื่องจาก Cox & Snell 's R^2 ในทางทฤษฎีไม่มีทางที่จะมีค่าสูงสุดเป็น 1 ดังนั้น Nagelkerke จึงได้มีการปรับปรุงสูตรของ Cox & Snell เล็กน้อย R^2 ค่าใหม่นี้สามารถคำนวณได้โดยใช้สูตรต่อไปนี้

Nagelkerke: $R^2_N = \frac{R^2_{CS}}{1 - e^{2LL(Baseline)/n}}$

Hosmer-Lemeshow goodness-of-fit test: มีการจัดแบ่งกลุ่มออกตาม predicted probabilities จากนั้นจะมีการคำนวณค่าของสถิติ χ^2 ตามสูตรดังนี้

$$\chi^2 = \Sigma (O-E)^2 / E$$ โดยใช้ degree of freedom = จำนวนกลุ่ม -2

โดยที่ χ^2 คือค่าสถิติไคว์สแควร์

O = observed frequency

E = expected frequency

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)