ตัวแบบถดถอยลอจิสติกอย่างง่าย

อรุณ จิรวัฒน์กุล*

ตัวแบบถดถอยลอจิสติก เป็นสมการถดถอยรูปแบบหนึ่งที่ใช้แสดงความสัมพันธ์ระหว่างตัวแปรตามที่ค่าตัวแปรจัดเป็น ๒ กลุ่ม (dichotomy) เช่น เป็นโรค/ไม่เป็นโรค, หาย/ตาย กับตัวแปรอิสระได้ทุกประเภท ทั้งตัวแปรนามสเกลที่มีค่าจัดได้ ๒ กลุ่ม หรือหลายกลุ่ม, ตัวแปรอันดับ, และตัวแปรต่อเนื่อง.

ตัวแบบถดถอยลอจิสติกนอกจากจะใช้ในการทำนายค่าตัวแปรตามแล้ว ค่าสัมประสิทธิ์ของตัวแบบยังสามารถคำนวณค่า Odds ratio ของการเกิดผล (ตัวแปรตาม) ที่มีอิทธิพลจากปัจจัยตัวแปรอิสระได้อีกด้วย. ในบทความนี้จะอธิบายแนวคิด, วิธีการสร้างตัวแบบ, การทดสอบตัวแบบ, และการคำนวณค่า Odd ratio จากตัวแบบ.

ตารางที่ ๑ ความสัมพันธ์ระหว่างอายุกับความเสี่ยงการเป็นโรคไม่ติดต่อ (NCD) ในตัวอย่าง ๑๒๐ คน

อายุ	คน	การเป็นโรคในกลุ่ม NCD	
		ราย	ความน่าจะเป็น
๔๒	๓๐	๑	๐.๐๓๓
๕๑	๒๒	๑	๐.๐๔๕
๖๐	๑๔	๖	๐.๔๒๙
๖๖	๑๗	๑๐	๐.๕๘๘
๗๒	๙	๖	๐.๖๖๗
๗๔	๑๖	๑๓	๐.๘๑๓
๘๕	๑๒	๑๐	๐.๘๓๓
รวม	๑๒๐	๔๗	

ทำไมจึงต้องใช้ตัวแบบถดถอยลอจิสติกกับตัวแปรตามที่ค่าตัวแปรจัดได้ ๒ กลุ่ม.

จากข้อมูลในตารางที่ ๑ แสดงว่าความน่าจะเป็นของการเป็นโรค NCD เพิ่มขึ้นตามอายุที่เพิ่มขึ้น. เมื่อนำข้อมูลมาสร้างแผนภาพขจายและสมการถดถอยเส้นตรง (รูปที่ ๑) พบว่าค่าข้อมูลที่แกน Y มีค่าเป็น ๐ หรือ ๑ เท่านั้น ทำให้ไม่มีค่าข้อมูลจริงอยู่บนเส้นกราฟเลย ซึ่งแสดงว่าการเป็นโรคกลุ่ม NCD ไม่มีความสัมพันธ์เชิงเส้นตรงกับอายุ.

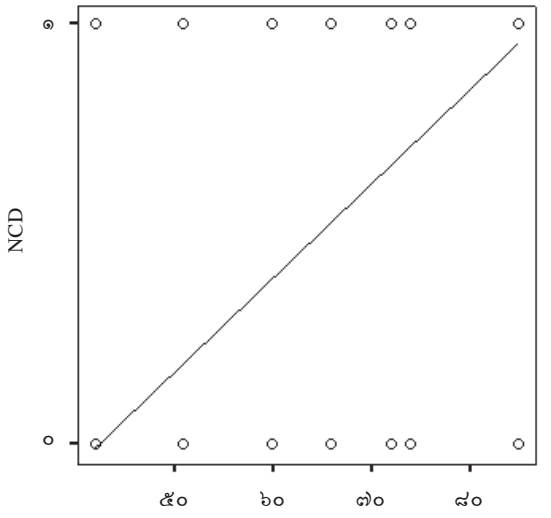
ดังนั้น จึงไม่สามารถใช้สมการเส้นตรงมาทำนายการเกิดโรคกลุ่ม NCD (Y) กับอายุ เพราะนอกจากข้อมูลไม่ได้มีความสัมพันธ์เชิงเส้นตรงแล้ว รูปแบบฟังค์ชั่นของสมการยังมีโอกาสที่ค่า Y จากการคำนวณมีค่า เกิน ๑ หรือต่ำกว่า ๐ ซึ่งเกินค่าข้อมูลจริงที่มีเฉพาะ ๐ กับ ๑ เท่านั้น.

ถ้าให้ P เป็นความน่าจะเป็นของการโรคกลุ่ม NCD (Y) นำเอาค่าความน่าจะเป็นของการเป็นโรคกลุ่ม NCD มาสร้างกราฟ กับอายุจะได้กราฟ รูปตัว S (รูปที่ ๒).

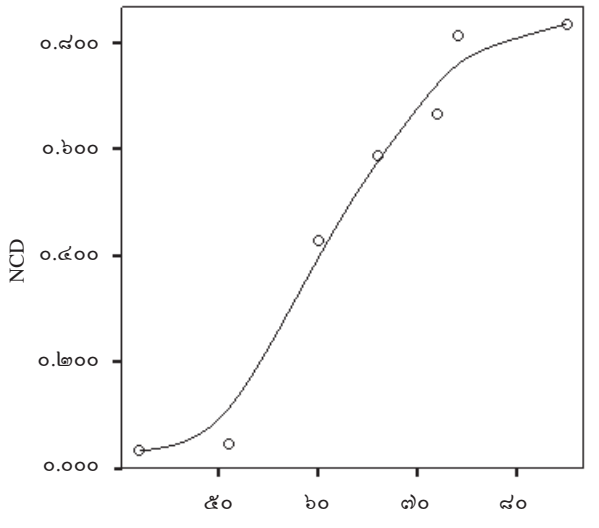
การสร้างสมการถดถอยลอจิสติก

ฟังค์ชั่นที่จะทำให้ความน่าจะเป็น (P) ที่คำนวณได้จากสมการมีค่าอยู่ระหว่าง ๐ กับ ๑ คือ

*ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น



รูปที่ ๑ แผนการขจายและสมการเส้นตรงระหว่างการเป็นโรคกลุ่ม NCD กับอายุของ



รูปที่ ๒ ความสัมพันธ์ระหว่างความน่าจะเป็นของการเป็นโรคกลุ่ม NCD กับอายุ

$$P = \frac{e^{a+bx}}{1 + e^{a+bx}} \quad \text{หรือ} \quad P = \frac{1}{1 + e^{a+bx}}$$

โดยที่ e = ค่า natural logarithm (e = ๒.๗๑๘)

a, b = ค่าสัมประสิทธิ์

x = ค่าตัวแปรอิสระ

ในการสร้างตัวแบบถดถอยลอจิสติก จะใช้ฟังค์ชั่นโลกิต $[Logit(P) = \ln\left\{\frac{P}{1 - P}\right\}]$ แปลงความสัมพันธ์ของความน่าจะเป็นของการเกิดโรค กับตัวแปรอิสระให้อยู่ในรูปความสัมพันธ์เชิงเส้นตรง การแปลงทำโดยการเปลี่ยนค่าความน่าจะเป็น (P) ให้อยู่ในรูปของค่า Odds $\left(\frac{P}{1 - P}\right)$ ซึ่งค่า log(Odds) ที่ได้จะมีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ (อายุ).

$$\text{Odds ของการเกิดโรค} = \frac{\text{ความน่าจะเป็นของการเกิดโรค}}{1 - \text{ความน่าจะเป็นของการเกิดโรค}}$$

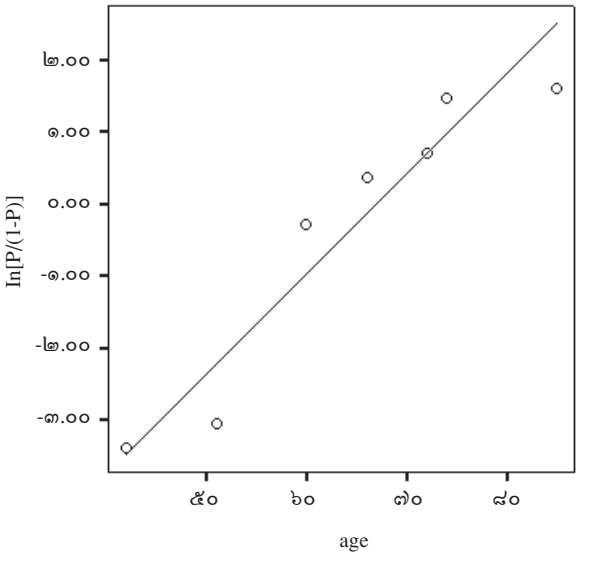
$$= \frac{P}{1 - P} \quad \text{แทนค่า} \quad P = \left\{\frac{e^{a+bx}}{1 + e^{a+bx}}\right\}$$

$$= e^{a+bx}$$

เมื่อทำให้เป็น log ฐาน e (ln) จะได้สมการของ log(Odds) ดังนี้

$$\ln\left\{\frac{P}{1 - P}\right\} = a + bx$$

จะได้ว่าค่า log(Odds) ของการเกิดโรคจะมีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ (X) เมื่อนำค่า log(Odds)



รูปที่ ๓ ความสัมพันธ์ระหว่าง log(Odds) ของการเป็นโรคกลุ่ม NCD กับอายุ

ตารางที่ ๒ ผลการคำนวณค่าพารามิเตอร์ a และ b ของตัวแบบถดถอยลอจิสติก

ตัวแปร	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% CI ของ EXP(B)	
							Lower	Upper
อายุ	๐.๑๓๗	๐.๐๒๔	๓๒.๔๕๓	๑	.๐๐๐	๑.๑๔๗	๑.๐๙๔	๑.๒๐๓
ค่าคงที่	-๘.๙๕๘	๑.๕๖๓	๓๓.๐๖๕	๑	.๐๐๐	.๐๐๐		

สร้างกราฟแสดงความสัมพันธ์กับอายุจะได้กราฟเส้นตรงดังแสดงในรูปที่ ๓.

การคำนวณค่าสัมประสิทธิ์ของตัวแบบ (a และ b)

การประมาณค่าสัมประสิทธิ์ a และ b ที่ทำให้ตัวแปรอิสระสามารถทำนายค่า Odds ของตัวแปรตามได้ดีที่สุดด้วยวิธี Maximum likelihood estimates (likelihood คือความน่าจะเป็นของ y เมื่อกำหนด x) ในการคำนวณจะเปลี่ยนค่าสัมประสิทธิ์ a และ b ไปที่ละค่า เพื่อหาค่าที่ทำให้ Log Likelihood (LL) ของสมการมีค่าสูงสุด. ค่า LL เป็นค่าที่ใช้แสดงว่าตัวแปรอิสระใช้ทำนายตัวแปรตามได้หรือไม่ ถ้าค่า LL มีค่าน้อย ๆ แสดงว่าใช้ทำนายตัวแปรตามได้ดี แต่ค่า LL ที่คำนวณได้มีค่าเป็นลบทำให้การเปรียบเทียบค่าทำได้ยาก จึงนำค่า -2 มาคูณ LL เพื่อให้ -2LL มีค่าเป็นบวก.

ดังนั้นในการพิจารณาว่าตัวแปรอิสระในสมการใช้ทำนายตัวแปรตามได้หรือไม่ จะพิจารณาจากค่า -2LL ถ้า -2LL มีค่าน้อย ๆ แสดงว่าตัวแปรอิสระใช้ทำนายตัวแปรตามได้ดี. ถ้า -2LL มีค่ามาก แสดงว่าตัวแปรอิสระนั้นไม่มีความสัมพันธ์กับตัวแปรตาม ไม่สามารถใช้เป็นตัวทำนาย.

การทดสอบว่าตัวแปรอิสระในตัวแบบใช้ทำนายตัวแปรตามหรือไม่

ในการพิจารณาว่าตัวแปรอิสระแต่ละตัวใช้ทำนายตัวแปรตามหรือไม่ ทำโดยใช้ Wald statistic หรือ $\text{Wald } \chi^2 = \left\{\frac{coefficient}{\text{SE coefficient}}\right\}^2$ ทดสอบว่าค่าสัมประสิทธิ์ของตัวแปรนั้นเท่ากับศูนย์หรือไม่ Wald statistic มีการแจกแจงแบบไคสแควร์ชั้นความเป็นอิสระ ๑ จากข้อมูลในตารางที่ ๑. ผลการทดสอบตัวแบบถดถอยลอจิสติก แสดงในตารางที่ ๒.

ค่าคงที่ คือค่า a พบว่า P value ของ Wald statistic น้อยกว่า ๐.๐๐๐๑ แสดงว่าค่าสัมประสิทธิ์ a มีค่าต่างจากศูนย์ที่ระดับนัยสำคัญ ๐.๐๕. โดยปรกติค่า a ไม่ใช่ค่าที่ใช้แสดงความสัมพันธ์ของตัวแปร จะต่างจากศูนย์ หรือไม่ต่างจากศูนย์ ก็ต้องมีไว้ในสมการ.

ส่วนค่า Wald statistic ของอายุ (b) มีค่า พี น้อยกว่า ๐.๐๐๐๑ สรุปได้ว่าค่าสัมประสิทธิ์ b มีค่าแตกต่างจากศูนย์อย่างมีนัยสำคัญ ๐.๐๕ แสดงว่าอายุใช้ทำนายการเป็นโรคกลุ่ม NCD ได้.

ค่าสัมประสิทธิ์ b กับ Odds ratio

Odds ของการเป็นโรค = e^{a+bx} หรือ = $e^a \cdot e^{bx}$ ถ้าค่าของตัวแปรอิสระเพิ่มขึ้น ๑ หน่วยจาก x เป็น x + ๑ ค่า Odds จะเพิ่มขึ้น จาก $e^a \cdot e^{bx}$ เป็น $e^a \cdot e^{b(x+1)}$ หรือ $e^a \cdot e^{bx} \cdot e^b$

$$\text{ดังนั้น Odds ratio (OR)} = \frac{e^a \cdot e^{bx} \cdot e^b}{e^a \cdot e^{bx}} e^b$$

$$\text{OR} = e^b$$

จะเห็นว่าเมื่อค่า x เพิ่มขึ้นหนึ่งหน่วย ค่า Odds ratio จะเพิ่ม e^b เท่า. ดังนั้นจึงสามารถคำนวณค่า Odds Ratio จากค่าสัมประสิทธิ์ b ของตัวแบบถดถอยลอจิสติกได้. จากตารางที่ ๒ ค่า b = ๐.๑๓๗ จะได้ค่า $R = 2.718^{0.137} = 1.147$.

ในการพิจารณาว่าค่า Odds ratio แตกต่างจาก ๑ หรือไม่ ให้พิจารณาจากช่วงเชื่อมั่น. จากตารางที่ ๒ พบว่าค่า ๙๕% ช่วงเชื่อมั่น Odds ของอายุ (CI ของ EXP(B)) อยู่ระหว่าง ๑.๐๙๔

ถึง ๑.๒๐๓ แสดงว่าปัจจัยอายุมีค่า Odds ratio เกิน ๑. เนื่องจากอายุที่ใช้คำนวณเป็นตัวแปรต่อเนื่อง ค่า Odds ratio ที่ได้จะเพิ่มขึ้นทุก ๆ อายุ ๑ ปี จึงทำให้ได้ค่า Odds ratio ค่อนข้างต่ำ เพราะในความเป็นจริงความเสี่ยงของการเป็นโรคในกลุ่ม NCD ไม่ได้เพิ่มขึ้นทุกปีตามอายุที่เพิ่มขึ้น แต่ความเสี่ยงจะเพิ่มขึ้นที่ละช่วงอายุ (อายุช่วงเดียวกันเสี่ยงเท่ากัน). ดังนั้นถ้าใช้หลักวิชาจัดกลุ่ม (ช่วง) อายุ ตามโอกาสเสี่ยงของการเป็นโรค จะทำให้กลุ่มอายุมีค่า Odds ratio สูงมากขึ้น.

ในกรณีตัวแปรตามที่ค่าตัวแปรจัดได้ ๒ กลุ่ม ค่า Odds ratio ที่คำนวณได้จากตัวแบบลอจิสติก จะมีค่าเท่ากับ Odds ratio ที่คำนวณจากตาราง ๒ x ๒ จากสูตร ad/bc.

สรุป

สมการถดถอยลอจิสติกใช้สำหรับสร้างตัวแบบการทำนายตัวแปรตามที่ค่าตัวแปรจัดได้ ๒ กลุ่ม ส่วนตัวแปรอิสระจะเป็นตัวแปรประเภทใดก็ได้. ค่าสัมประสิทธิ์ของตัวแปรอิสระ b ใช้คำนวณค่า Odds ratio ได้.

ในการสร้างตัวแบบเพื่อทำนายผลจะต้องพิจารณาความเหมาะเจาะ (fitness) ของสมการว่าจะใช้ทำนายได้ดีหรือไม่. การทดสอบความเหมาะเจาะและตัวแบบพหุที่มีตัวแปรอิสระมากกว่า ๑ ตัว จะได้อธิบายในบทความต่อไป.

เอกสารประกอบการเรียบเรียง

๑. อรุณ จิรวัฒน์กุล. ชีวสถิติที่ใช้ในงานวิจัยทางวิทยาศาสตร์สุขภาพ. พิมพ์ครั้งที่ ๑. ขอนแก่น : คณะสาธารณสุขศาสตร์; ๒๕๐๔.

๒. Kleinbaum DG, Klein M. Logistic regression: a self-learning text. New York: Springer; 2002.