

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การถดถอย (regression analysis) เป็นวิธีการทางสถิติอย่างหนึ่งที่ใช้ศึกษาถึงลักษณะความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยพิจารณาถึงรูปแบบความเป็นไปได้ของความสัมพันธ์ที่มีเป้าหมายของการวิเคราะห์คือหาสมการทางคณิตศาสตร์ที่ใช้เป็นตัวแทนแสดงความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยมีวัตถุประสงค์ที่จะประมาณค่าหรือพยากรณ์ค่าของตัวแปรหนึ่งจากค่าที่กำหนดให้ของตัวแปรอื่นๆ ที่เกี่ยวข้องกัน ตัวแปรในตัวแบบการถดถอยนี้สามารถแบ่งออกเป็น 2 ประเภท ประเภทแรกคือตัวแปรที่ต้องการประมาณหรือพยากรณ์เรียกว่าตัวแปรตาม (dependent variable) หรือตัวแปรตอบสนอง (response variable) ตัวแปรประเภทที่สองคือตัวแปรที่มีอิทธิพลต่อตัวแปรตามหรือตัวแปรตอบสนองเรียกว่าตัวแปรอธิบาย (explain variable) หรือตัวแปรอิสระ (independent variable) การวิเคราะห์การถดถอยประกอบด้วยขั้นตอนต่างๆ หลายขั้นตอน และขั้นตอนที่สำคัญขั้นตอนหนึ่งก็คือการคัดเลือกตัวแปรอธิบายที่มีอิทธิพลต่อการอธิบายความแปรผันของตัวแปรตาม ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson correlation coefficient) เป็นค่าเบื้องต้นที่ใช้วัดความสัมพันธ์ในลักษณะเชิงเส้นระหว่างตัวแปรอธิบายและตัวแปรตาม รวมทั้งการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยนั้นก็เป็นขั้นตอนที่สำคัญขั้นตอนหนึ่ง การประมาณค่าพารามิเตอร์มีหลายวิธีด้วยกันแต่วิธีที่นิยมใช้กันมากได้แก่วิธีกำลังสองน้อยสุด (least squares method) และวิธีความควรจะเป็นสูงสุด (maximum likelihood method)

เมื่อได้ตัวแบบการถดถอยแล้ว จะต้องทำการทดสอบนัยสำคัญของสัมประสิทธิ์การถดถอย เพื่อให้ประกอบการตัดสินใจว่าตัวแปรอธิบายตัวใดบ้างควรอยู่ในตัวแบบการถดถอย และตัวแปรอธิบายตัวใดบ้างที่มีอิทธิพลต่อการอธิบายความแปรผันของตัวแปรตาม สำหรับเกณฑ์ในการตัดสินใจว่าตัวแบบการถดถอยที่ได้นั้นเหมาะสมหรือไม่ มีค่าสถิติหลายตัวที่ช่วยในการตัดสินใจ เช่นอาจพิจารณาได้จากค่าสัมประสิทธิ์การตัดสินใจ (coefficient of determination) ซึ่งเป็นค่าที่นิยมใช้กันมากในปัจจุบันเนื่องจากง่ายในการตีความ และเป็นค่าที่คำนวณจากสัดส่วน

ความแปรผันของตัวแปรตามที่ตัวแบบการถดถอยสามารถอธิบายได้ อย่างไรก็ตามถึงแม้ว่าสัมประสิทธิ์การตัดสินใจเป็นค่าที่นิยมใช้ในการวัดความเหมาะสมของตัวแบบการถดถอย แต่สำหรับการวิเคราะห์การถดถอยที่ประกอบด้วยตัวแปรอธิบายที่มีมากกว่า 1 ตัว เกิดปัญหาที่สำคัญอย่างหนึ่ง คือบางครั้งสัมประสิทธิ์การตัดสินใจให้ขนาดของความสัมพันธ์ที่มากเกินไป ทั้งที่ตัวแปรอธิบายนั้นมีความสัมพันธ์กับตัวแปรตามน้อยมาก และในความเป็นจริงพบว่าการเพิ่มตัวแปรอธิบายเข้าไปในตัวแบบการถดถอยนั้น ค่าสัมประสิทธิ์การตัดสินใจจะไม่ลดลง แม้ว่าตัวแปรอธิบายที่เพิ่มเข้าไปจะเป็นอิสระกับตัวแปรตามก็ตาม ดังนั้นค่าของสัมประสิทธิ์การตัดสินใจในบางตัวแบบจะมีค่าสูงขึ้นเมื่อจำนวนตัวแปรอธิบายเพิ่มขึ้น และจะส่งผลให้ค่าสัมประสิทธิ์การตัดสินใจของตัวแบบนี้เข้าใกล้หนึ่งมากขึ้น สัมประสิทธิ์การตัดสินใจที่ปรับค่าจะแก้ปัญหานี้ได้ ด้วยการให้จำนวนตัวแปรอธิบายที่มีในตัวแบบมาช่วยในการคำนวณ และถือว่าเป็นค่าที่ดีและมีประโยชน์สำหรับใช้ในการเปรียบเทียบตัวแบบการถดถอยสองตัวแบบที่มีจำนวนตัวแปรอธิบายที่ต่างกัน

ในการวิเคราะห์การถดถอย จะต้องทำภายใต้ข้อสมมุติ (assumptions) เกี่ยวกับค่าคลาดเคลื่อนสุ่ม (random error) ดังนั้นเมื่อได้ตัวแบบการถดถอยแล้วจึงจำเป็นต้องอย่างยิ่งที่จะต้องตรวจสอบค่าคลาดเคลื่อนสุ่มว่าสอดคล้องกับข้อสมมุติหรือไม่ เนื่องจากไม่อาจทราบค่าคลาดเคลื่อนสุ่มจริงได้ ดังนั้นจึงใช้ส่วนตกค้าง (residuals) ที่เป็นค่าประมาณของค่าคลาดเคลื่อนสุ่มมาตรวจสอบแทน บ่อยครั้งที่ผลการตรวจสอบเหล่านี้ถูกนำไปใช้ในการพิจารณาปรับแก้ตัวแบบการถดถอยใหม่ เพื่อให้ได้ตัวแบบที่เหมาะสมยิ่งขึ้น ตัวแบบการถดถอยเชิงเส้นมีรูปแบบดังนี้

$$Y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip} + e_i \quad ; i = 1, \dots, n$$

เมื่อ	Y_i	คือตัวแปรตามซึ่งเป็นตัวแปรสุ่ม
	$x_{i1}, x_{i2}, \dots, x_{ip}$	คือค่าของตัวแปรอธิบายและมีค่าคงที่ไม่ใช่ค่าสุ่ม
	$b_0, b_1, b_2, \dots, b_p$	คือพารามิเตอร์ของตัวแบบ
	e_i	คือค่าคลาดเคลื่อนสุ่ม

โดยมีข้อสมมุติในการวิเคราะห์ตัวแบบการถดถอยดังนี้

1. e_i เป็นตัวแปรสุ่ม
2. e_i มีการแจกแจงปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่ (s^2)

3. e_i และ e_j เป็นอิสระซึ่งกันและกัน นั่นคือ $Cov(e_i, e_j) = 0$ โดยที่ $i \neq j$
4. Y_i และ Y_j จะต้องเป็นอิสระซึ่งกันและกัน และมีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ $b_0 + b_1x_{i1} + \dots + b_px_{ip}$ และความแปรปรวนคงที่ (s^2)

จากที่ได้กล่าวไปแล้วนั้นเป็นการวิเคราะห์การถดถอยเชิงเส้นที่ตัวแปรตามเป็นตัวแปรเชิงปริมาณ แต่ถ้าตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่จำแนกข้อมูลออกเป็นสองกลุ่มเรียกว่าตัวแปรทวิภาค (binary responses variable) โดยที่กำหนดให้มีค่าเป็น 1 เมื่อตัวแปรตามเกิดลักษณะที่สนใจและมีค่าเป็น 0 ในกรณีอื่นๆ นั้นจะใช้วิธีการทางสถิติที่เรียกว่าการวิเคราะห์การถดถอยโลจิสติกทวิภาค (binary logistic regression analysis) หรือเรียกย่อๆ ว่าการวิเคราะห์การถดถอยโลจิสติก เพราะเป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตามเชิงคุณภาพและตัวแปรอธิบายอื่นๆ โดยที่ตัวแปรอธิบายอาจเป็นตัวแปรเชิงคุณภาพหรือเชิงปริมาณก็ได้ เพื่อนำมาใช้ในการพยากรณ์ค่าความน่าจะเป็นของสิ่งที่เราสนใจ ความสัมพันธ์ระหว่างความน่าจะเป็นที่ตัวแปรตามเกิดลักษณะที่สนใจกับตัวแปรอธิบายเหล่านี้สามารถแสดงในรูปของตัวแบบการถดถอยโลจิสติกได้ดังนี้

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = b_0 + \sum_{j=1}^p b_j x_{ij} \quad ; i=1, \dots, n, j=1, \dots, p$$

เมื่อ	$p_i = P(Y_i = 1 X = x_i)$	แทนความน่าจะเป็นแบบมีเงื่อนไขที่ตัวแปรตามเกิดลักษณะที่สนใจ
	b_0, b_j	คือพารามิเตอร์ของตัวแบบ
	Y_i	คือตัวแปรตามแบบทวิภาค
	x_{ij}	คือค่าของตัวแปรอธิบาย

สำหรับการกำหนดตัวแบบการถดถอยโลจิสติกนี้ เราต้องเลือกตัวแบบที่สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอธิบายให้สมเหตุสมผล โดยการเลือกตัวแบบการถดถอยโลจิสติกนั้นจะต้องเลือกตัวแบบให้เหมาะสมกับข้อมูลให้มากที่สุด โดยเฉพาะข้อมูลเชิงคุณภาพที่อยู่ในรูปตารางการถ่วงน้ำหนักของจำนวนนับต่างๆ

การตรวจสอบตัวแบบการถดถอยโลจิสติกนี้ สามารถทำการตรวจสอบภายใต้เงื่อนไขต่างๆ ได้ อาจตรวจสอบโดยใช้ตัวสถิติอัตราส่วนของฟังก์ชันควรจะเป็นแบบมีเงื่อนไขสำหรับการเปรียบเทียบของ 2 ตัวแบบ โดยที่ตัวแบบหนึ่งเป็นกรณีพิเศษของอีกตัวแบบหนึ่ง สำหรับตัวสถิติอัตราส่วนของฟังก์ชันควรจะเป็นแบบมีเงื่อนไขนี้ จะมีการแจกแจงเมื่อใกล้อนันต์แบบไคกำลังสอง (asymptotic chi-square distribution) และมีองศาเสรี (degrees of freedom) เท่ากับค่าแตกต่างระหว่างองศาเสรีของตัวแบบทั้งสอง นอกจากนี้การตรวจสอบความเหมาะสมของตัวแบบนั้นอาจตรวจสอบได้โดยใช้สัมประสิทธิ์การตัดสินใจ (R^2) การคำนวณค่า R^2 สำหรับตัวแบบการถดถอยโลจิสติกนั้นทำได้หลายวิธี เช่นวิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least squares, OLS) และวิธีความควรจะเป็นสูงสุด (maximum likelihood, ML) ได้มีผู้ศึกษาค่า R^2 ของการถดถอยโลจิสติกไว้หลายท่าน เช่น มิทเบ็คและสเชมเปอร์ (Mittlbach and Schemper, 1996) อกรีติ (Agresti, 2000) และมินาร์ด (Menard, 2000) และผลการศึกษาพบว่า R^2 ในการถดถอยโลจิสติกจะเหมือนกับค่า R^2 ในการถดถอยเชิงเส้น ที่บางครั้งก็ให้ขนาดของความสัมพันธ์มากเกินไป แม้ว่าตัวแปรอธิบายที่เพิ่มเข้าไปในตัวแบบนั้นจะไม่มีความสัมพันธ์กับตัวแปรตาม ดังนั้นสำหรับตัวแบบการถดถอยโลจิสติกที่ประกอบด้วยตัวแปรอธิบายที่มีมากกว่าหนึ่งตัวแปร จึงจำเป็นต้องมีการปรับค่า R^2

การปรับค่าสัมประสิทธิ์การตัดสินใจนั้น โดยส่วนใหญ่จะใช้วิธีการปรับ 2 แบบด้วยกัน แบบแรกจะใช้จำนวนตัวแปรอธิบายที่มีในตัวแบบและขนาดตัวอย่าง มาช่วยในการคำนวณ หรือเรียกว่าการใช้องศาเสรี (degrees of freedom) และแบบที่สองใช้ผลต่างของค่าฟังก์ชันควรจะเป็นของตัวแบบ 2 ตัวแบบหรือที่เรียกว่าสถิติดีเวียนซ์ (deviance) ในปี ค.ศ. 1996 มิทเบ็คและสเชมเปอร์ (Mittlbach and Schemper) ได้ศึกษาค่า R^2 ที่ปรับค่าโดยใช้รูปแบบทั้งสอง โดยในที่นี้ใช้สัญลักษณ์ $R_{O,adj,MS}^2$ และ $R_{l,adj,MS}^2$ สำหรับการปรับค่าที่ใช้องศาเสรีและสถิติดีเวียนซ์ ตามลำดับ ผลการศึกษาพบว่าค่า R_{adj}^2 ทั้งสองตัวนี้ให้ผลเป็นที่น่าพอใจกว่าค่า R^2 ที่ยังไม่ได้ทำการปรับค่า รูปแบบของ $R_{O,adj,MS}^2$ และ $R_{l,adj,MS}^2$ แสดงได้ดังนี้

$$R_{O,adj,MS}^2 = 1 - \frac{(n-1) \sum_{i=1}^n (y_i - \hat{p}_i)^2}{(n-p-1) \sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.1)$$

$$R_{l,adj,MS}^2 = 1 - \frac{l(y, \hat{p}) - (p+1)/2}{l(y, \hat{p}^0) - 1/2} \quad (1.2)$$

เมื่อ	y_i	คือค่าสังเกตของตัวแปรตามแบบทวิภาค
	n	คือขนาดตัวอย่าง
	p	คือจำนวนตัวแปรอธิบายในตัวแบบการถดถอย
	$\hat{p}_i$	คือค่าประมาณของ p_i กรณีที่ตัวแบบมีตัวแปรอธิบาย p ตัว
	$\bar{y}$	คือค่าเฉลี่ยของตัวแปรตามแบบทวิภาค $\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$
	$l(y, \hat{p})$	คือฟังก์ชันลอกลิควรรจะเป็นสูงสุดสำหรับตัวแบบที่มีตัวแปรอธิบาย p ตัว ซึ่งคำนวณจาก $\sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$
	$l(y, \hat{p}^0)$	คือฟังก์ชันลอกลิควรรจะเป็นสูงสุดสำหรับตัวแบบที่มีเพียงเทอมค่าคงที่ ซึ่งคำนวณจาก $\sum_{i=1}^n [y_i \log(\hat{p}_i^0) + (1 - y_i) \log(1 - \hat{p}_i^0)]$; $\hat{p}_i^0 = \bar{y}$

ในปี ค.ศ. 2003 เลียโอและแมคกี (Liao and McGee) ได้เสนอ R_{adj}^2 ของการถดถอยโลจิสติกเพื่อใช้แก้ปัญหาของ R^2 ที่ให้ขนาดความสัมพันธ์ที่มากเกินไป และมีค่าสูงขึ้นเมื่อขนาดตัวอย่างลดลง โดยที่ตัวแบบมีตัวแปรอธิบายมากใกล้ตัวแบบเต็ม (saturated model) โดยมีการปรับค่า R^2 ด้วยค่าคลาดเคลื่อนที่คำนวณโดยใช้ขั้นตอน 2 ขั้นตอนคือ ขั้นที่ 1 คำนวณค่าคลาดเคลื่อนของการพยากรณ์ โดยจะเรียกค่าคลาดเคลื่อนนี้ว่าค่าคลาดเคลื่อนของการพยากรณ์ที่ไม่สามารถแยกได้ (inherent prediction error , IPE) ของตัวแบบโลจิสติก ขั้นที่ 2 ปรับค่าคลาดเคลื่อนของการพยากรณ์ที่ได้ในขั้นที่ 1 โดยใช้หลักการเช่นเดียวกับการประมาณค่าความแปรปรวน (s^2) ในตัวแบบการถดถอยด้วยวิธี REML (restricted or residual maximum likelihood) โดยที่ $\hat{IPE}_0^p$ และ $\hat{IPE}_0^0$ คือค่าประมาณของ IPE สำหรับตัวแบบที่มีตัวแปรอธิบาย p ตัว และตัวแบบที่มีเพียงเทอมของค่าคงที่ ตามลำดับ เมื่อคำนวณโดยวิธีกำลังสองน้อยสุดแบบสามัญ (OLS) และ $\hat{IPE}_1^p$ และ $\hat{IPE}_1^0$ คือค่าประมาณของ IPE สำหรับตัวแบบที่มีตัวแปรอธิบาย p ตัว และตัวแบบที่มีเพียงเทอมของค่าคงที่ ตามลำดับ เมื่อคำนวณโดยวิธีความควรจะเป็นสูงสุด (ML) ผลการศึกษาพบว่าเมื่อเพิ่มตัวแปรอธิบายที่เป็นอิสระกับตัวแปรตามเข้าในตัวแบบค่า $R_{l,adj,LM}^2$ ให้ผลที่ดีกว่าค่า $R_{O,adj,MS}^2$ โดยเฉพาะตัวแบบที่

มีขนาดตัวอย่างเล็กหรือตัวแบบที่มีจำนวนตัวแปรอธิบายมาก รูปแบบ $R_{O,adj,LM}^2$ และ $R_{I,adj,LM}^2$ แสดงได้ดังนี้

$$R_{O,adj,LM}^2 = 1 - \frac{\hat{IPE}_O^p}{\hat{IPE}_O^o} \quad (1.3)$$

$$R_{I,adj,LM}^2 = 1 - \frac{\hat{IPE}_I^p}{\hat{IPE}_I^o} \quad (1.4)$$

ในปัจจุบันโปรแกรมสำเร็จรูปทางสถิติได้ให้ผลลัพธ์ของค่า R^2 ในการวิเคราะห์การถดถอยโลจิสติก เช่นโปรแกรม SAS[®] ในคำสั่ง PROC LOGISTIC และ PROC GENMOD ได้รายงานตัววัด R^2 หลายตัว (Shtatland, 2000) แต่ที่นิยมใช้กันมากได้แก่ $R_{I,adj,SAS_{DEV}}^2$ และ $R_{I,adj,SAS_{AIC}}^2$ ที่มีรูปแบบดังนี้

$$R_{I,adj,SAS_{DEV}}^2 = \frac{1 - \exp\{-2[l(y, \hat{p}) - l(y, \hat{p}^0)]/n\}}{1 - \exp[2l(y, \hat{p}^0)/n]} \quad (1.5)$$

$$R_{I,adj,SAS_{AIC}}^2 = 1 - \frac{l(y, \hat{p}) - [(p+1)(n-1)/(n-p-1)]}{l(y, \hat{p}^0) - 1} \quad (1.6)$$

เมื่อ $l(y, \hat{p})$ $l(y, \hat{p}^0)$ n และ p นิยามได้เช่นเดียวกับสมการ (1.2) ตัววัด $R_{I,adj,SAS_{DEV}}^2$ ได้รับความนิยมน้อยมากในโปรแกรม SAS[®] แต่ไม่ได้หมายความว่าการใช้เพียงค่า $R_{I,adj,SAS_{DEV}}^2$ จะเพียงพอสำหรับการวัดความเหมาะสมของตัวแบบ เพราะในบางครั้ง $R_{I,adj,SAS_{DEV}}^2$ ให้ค่าที่สูงเกินไป ดังนั้นจึงควรใช้ควบคู่กับตัววัด $R_{I,adj,SAS_{AIC}}^2$ ที่ได้จากการปรับค่าตามหลักเกณฑ์สารสนเทศของอาไคเคะ (Akaike's information criterion, AIC)

จากงานวิจัยข้างต้นได้มีการนำเสนอสมประสิทธิ์การตัดสินใจที่ปรับค่า (R_{adj}^2) ของตัวแบบการถดถอยโลจิสติกแบบต่างๆ เมื่อตัวแปรตามเป็นแบบทวิภาค โดยจะเห็นได้ว่าตัววัด R^2 เหล่านี้มีพื้นฐานการคำนวณโดยใช้หลักการที่สำคัญสองวิธี ได้แก่วิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least squares, OLS) และวิธีความควรจะเป็นสูงสุด (maximum likelihood, ML) และยังแสดงให้เห็นถึงประสิทธิภาพของ R_{adj}^2 แบบต่างๆ เช่น $R_{O,adj,MS}^2$ เหมาะสมสำหรับใช้

วัดความเหมาะสมของตัวแบบที่มีอัตราส่วน p/n สูง (ในงานที่ผ่านมามีค่าเท่ากับ 0.2) ในกรณีที่ตัวแปรอธิบายมีการแจกแจงเบร์นูลลี และค่า $R_{l,adj,LM}^2$ เหมาะสมสำหรับใช้วัดความเหมาะสมของตัวแบบกรณีที่ตัวแปรอธิบายมีการแจกแจงเอกรูปที่ต่อเนื่อง รวมถึงตัววัด $R_{l,adj,SAS_{DEV}}^2$ และ $R_{l,adj,SAS_{AIC}}^2$ ที่ทำงานได้อย่างถูกต้องตามหลักเกณฑ์สารสนเทศของอาไคเคะ (AIC) กรณีที่ขนาดตัวอย่างเล็ก

อย่างไรก็ตามยังไม่มียานวิจัยที่ศึกษาผลของการกำหนดตัวแบบการถดถอยไม่ถูกต้องที่มีต่อค่าสัมประสิทธิ์การตัดสินใจ (R^2) สำหรับตัวแบบการถดถอยโลจิสติก ดังนั้นในงานวิจัยนี้จึงสนใจศึกษาเปรียบเทียบประสิทธิภาพของสัมประสิทธิ์การตัดสินใจ R_O^2 , R_I^2 และสัมประสิทธิ์การตัดสินใจที่ปรับค่า $R_{O,adj,MS}^2$, $R_{I,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{I,adj,LM}^2$, $R_{l,adj,SAS_{DEV}}^2$ และ $R_{l,adj,SAS_{AIC}}^2$ ของตัวแบบการถดถอยโลจิสติก และศึกษาอิทธิพลของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อสัมประสิทธิ์การตัดสินใจและสัมประสิทธิ์การตัดสินใจที่ปรับค่าดังที่กล่าวข้างต้น โดยที่การกำหนดตัวแบบไม่ถูกต้องจะศึกษาใน 3 ลักษณะได้แก่ ลักษณะที่รูปแบบของตัวแปรอธิบายไม่ถูกต้อง (wrong functional form of the covariate) ลักษณะที่กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง (misspecified link function) และลักษณะที่กำหนดให้ตัวแบบมีตัวแปรอธิบายที่มีความสัมพันธ์กับตัวแปรตามขาดหายไป (missing covariate) สำหรับตัวแบบการถดถอยโลจิสติกที่ทำการศึกษานี้จะประกอบด้วยตัวแปรอธิบายจำนวน 1, 5 และ 10 ตัว ในกรณีที่ตัวแบบมีตัวแปรอธิบาย 1 ตัว กำหนดให้ตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณและตัวแบบที่มีตัวแปรอธิบายมากกว่า 1 ตัว จะกำหนดให้ตัวแปรอธิบายมีทั้งตัวแปรเชิงคุณภาพและเชิงปริมาณ โดยกำหนดให้ตัวแปรอธิบายที่อยู่ในตัวแบบทุกตัวเป็นตัวแปรที่มีความสัมพันธ์อย่างมีนัยสำคัญกับตัวแปรตาม

วัตถุประสงค์ของการวิจัย

1. เปรียบเทียบประสิทธิภาพของสัมประสิทธิ์การตัดสินใจ ของตัวแบบการถดถอยโลจิสติก ที่คำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least squares) R_O^2 , $R_{O,adj,MS}^2$, $R_{O,adj,LM}^2$ และวิธีความควรจะเป็นสูงสุด (maximum likelihood) R_I^2 , $R_{I,adj,MS}^2$, $R_{I,adj,LM}^2$, $R_{l,adj,SAS_{DEV}}^2$, $R_{l,adj,SAS_{AIC}}^2$

2. ศึกษาความเอนเอียงสัมพัทธ์ (relative bias) และรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error, RMSE) ของสัมประสิทธิ์การตัดสินใจ (R_O^2 , R_I^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า ($R_{O,adj,MS}^2$, $R_{I,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{I,adj,LM}^2$, $R_{I,adj,SAS_{DEV}}^2$, $R_{I,adj,SAS_{AIC}}^2$) ของตัวแบบการถดถอยโลจิสติก เมื่อกำหนดตัวแบบไม่ถูกต้อง

สมมุติฐานของการวิจัย

1. สัมประสิทธิ์การตัดสินใจที่ปรับค่า (R_{adj}^2) ของตัวแบบการถดถอยโลจิสติกที่คำนวณโดยใช้หลักการเดียวกันมีประสิทธิภาพไม่แตกต่างกัน
2. สัมประสิทธิ์การตัดสินใจที่ปรับค่า (R_{adj}^2) ที่คำนวณโดยวิธีความควรจะเป็นสูงสุด ($R_{I,adj,MS}^2$, $R_{I,adj,LM}^2$, $R_{I,adj,SAS_{DEV}}^2$, $R_{I,adj,SAS_{AIC}}^2$) มีประสิทธิภาพมากกว่าวิธีกำลังสองน้อยสุดแบบสามัญ ($R_{O,adj,MS}^2$, $R_{O,adj,LM}^2$) สำหรับตัวแบบการถดถอยโลจิสติก
3. สัมประสิทธิ์การตัดสินใจ (R_O^2 , R_I^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า ($R_{O,adj,MS}^2$, $R_{I,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{I,adj,LM}^2$, $R_{I,adj,SAS_{DEV}}^2$, $R_{I,adj,SAS_{AIC}}^2$) ของตัวแบบการถดถอยโลจิสติกที่คำนวณโดยใช้หลักการเดียวกันมีความเอนเอียงสัมพัทธ์ และรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยไม่แตกต่างกัน เมื่อกำหนดตัวแบบไม่ถูกต้อง

ขอบเขตการศึกษา

งานวิจัยนี้เป็นการศึกษาโดยใช้การจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล ที่มีขอบเขตการศึกษาดังนี้

1. ตัวแปรที่ใช้ในการศึกษา
 - 1.1 การแจกแจงของตัวแปรอธิบาย X_j โดยที่ $j=1,2,...,10$ กำหนดให้เป็นตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ ดังนี้
 - X_1 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 2 และมีความแปรปรวนเท่ากับ 4 ($Normal(2,4)$)
 - X_2 มีการแจกแจงเอกรูปไม่ต่อเนื่องที่มีค่าเป็นไปได้อีกคือ 0 และ 1 ($Uniform[0,1]$)
 - X_3 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 20 และมีความแปรปรวนเท่ากับ 4 ($Normal(20,4)$)

- X_4 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1, 2 และ 3
 ($Uniform[0,3]$)
- X_5 มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 8 และมีความแปรปรวนเท่ากับ 3
 ($Normal(8,3)$)
- X_6 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1 และ 2
 ($Uniform[0,2]$)
- X_7 มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 0 และมีความแปรปรวนเท่ากับ 1
 ($Normal(0,1)$)
- X_8 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0 และ 1
 ($Uniform[0,1]$)
- X_9 มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 15 และมีความแปรปรวนเท่ากับ 2
 ($Normal(15,2)$)
- X_{10} มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1, 2, 3 และ 4
 ($Uniform[0,4]$)

1.2 กำหนดตัวแปรตามเป็นแบบทวิภาคที่มีการแจกแจงแบร์นูลลี โดยมีโอกาสเกิดความสำเร็จเท่ากับ p_i ($Y_i \sim Bernoulli(p_i)$) ในการจำลองชุดข้อมูลของตัวแปรตามจะจำลองจากตัวแบบ 3 ตัวแบบ ซึ่งสัมประสิทธิ์การถดถอยของตัวแบบจะถูกกำหนดขึ้นมาเพื่อให้ค่าของตัวแปรตามมีค่า 0 และ 1 ในสัดส่วนที่ใกล้เคียงกัน

$$\text{logit}(p_i) = b_0 + b_1 x_{i1}$$

$$\text{logit}(p_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5}$$

$$\text{logit}(p_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8} + b_9 x_{i9} + b_{10} x_{i10}$$

2. ตัวแบบที่ใช้ในการศึกษามี 3 ตัวแบบที่ประกอบด้วยตัวแปรอธิบายที่แตกต่างกัน

2.1 ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 1 ตัว กำหนดให้เป็นตัวแปรเชิงปริมาณที่ได้จากการแจกแจงปรกติ (X_1)

$$\text{logit}(p_i) = b_0 + b_1 x_{i1}$$

2.2 ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 5 ตัว โดยมีทั้งตัวแปรเชิงคุณภาพและเชิงปริมาณที่ได้จากการแจกแจงปกติ (X_1, X_3, X_5) และการแจกแจงเอกรูปไม่ต่อเนื่อง (X_2, X_4)

$$\text{logit}(p_i) = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5}$$

2.3 ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 10 ตัว โดยมีทั้งตัวแปรเชิงคุณภาพและเชิงปริมาณที่ได้จากการแจกแจงปกติ (X_1, X_3, X_5, X_7, X_9) และการแจกแจงเอกรูปไม่ต่อเนื่อง ($X_2, X_4, X_6, X_8, X_{10}$)

$$\text{logit}(p_i) = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6} + b_7x_{i7} + b_8x_{i8} + b_9x_{i9} + b_{10}x_{i10}$$

3. กำหนดขนาดตัวอย่าง (n) ที่ทำการศึกษาคือ 50, 100, 1000 และ 2000

4. กำหนดรูปแบบของตัวแบบการถดถอยโลจิสติกที่ไม่ถูกต้องใน 3 ลักษณะคือ

4.1 กำหนดรูปแบบของตัวแปรอธิบายไม่ถูกต้อง (wrong functional form of the covariate) โดยกำหนดให้ตัวแปรอธิบายที่เป็นตัวแปรเชิงปริมาณ (X_1) อยู่ในรูปกำลังสอง (X_1^2)

4.2 กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง (misspecified link function) โดยกำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้อง 2 รูปแบบดังนี้
รูปแบบแรกคือฟังก์ชันคอมพลีเมนทารีลอจิสติก (complementary log-log)

$$\log[-\log(1-p_i)] = b_0 + \sum_j b_j x_{ij} \quad ; i=1,2,\dots,n, j=1,2,\dots,p$$

รูปแบบที่สองคือฟังก์ชันโพรบิต (probit)

$$\Phi^{-1}(p_i) = b_0 + \sum_j b_j x_{ij} \quad ; i=1,2,\dots,n, j=1,2,\dots,p$$

4.3 กำหนดให้ตัวแบบการถดถอยมีตัวแปรอธิบายที่มีความสัมพันธ์กับตัวแปรตามขาดหายไป (missing covariate) ศึกษาเฉพาะตัวแบบที่มีจำนวนตัวแปรอธิบายเท่ากับ 5 และ 10 ตัว

5. การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยของตัวแบบการถดถอยโลจิสติกจะใช้วิธีความควรจะเป็นสูงสุดของฟิชเชอร์สกอริง (Fisher's scoring method) โดยใช้โปรแกรม SAS® เวอร์ชัน 8.0 ในการคำนวณ
6. ในการจำลองจะทำซ้ำจำนวน 1,000 ครั้ง ในแต่ละตัวแบบการถดถอย ภายใต้ขนาดตัวอย่างต่างๆ

เกณฑ์ที่ใช้ในการเปรียบเทียบ

เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของสัมประสิทธิ์การตัดสินใจ (R^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า (R_{adj}^2) มีดังต่อไปนี้

1. ค่ามัธยฐาน (median, $\hat{R}^2$) ของค่าประมาณสัมประสิทธิ์การตัดสินใจ ($R_{est,k}^2$) ที่คำนวณจากการทำซ้ำจำนวน 1,000 ครั้ง เมื่อ $k = 1, 2, \dots, 1000$ มีสูตรดังนี้

$$\hat{R}^2 = \frac{R_{est(500)}^2 + R_{est(501)}^2}{2}$$

โดยจะเปรียบเทียบความเอนเอียง (bias) ของค่าประมาณมัธยฐาน (median, $\hat{R}^2$) ที่แตกต่างจากค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง (R_{true}^2) และความเอนเอียงสัมพัทธ์ (relative bias) ของค่าประมาณมัธยฐาน ($\hat{R}^2$) ที่ได้จากการทำซ้ำจำนวน 1,000 ครั้ง โดยที่

$$\text{bias} = \hat{R}^2 - R_{true}^2 \quad \text{และ} \quad \text{relative bias} = \left| \frac{\hat{R}^2 - R_{true}^2}{R_{true}^2} \right| \times 100\%$$

เมื่อ $R_{est(500)}^2$ และ $R_{est(501)}^2$ คือค่าประมาณของ R^2 หรือ R_{adj}^2 ที่อยู่ในลำดับที่ 500 และ 501 จากการเรียงลำดับของค่าประมาณของ R^2 หรือ R_{adj}^2 ที่ได้จากการทำซ้ำจำนวน 1,000 ครั้ง

และ $R_{O,true}^2$ และ $R_{I,true}^2$ เป็นค่าสัมประสิทธิ์การตัดสินใจที่แท้จริงจากการคำนวณด้วยวิธีกำลังสองน้อยสุดและวิธีความควรจะเป็นสูงสุด ตามลำดับ โดยมีสูตรคำนวณดังนี้

$$R_{O,true}^2 = 1 - \frac{\sum_{i=1}^n [p_i(1-p_i)]}{\sum_{i=1}^n [\bar{p}(1-\bar{p})]} \quad (1.7)$$

$$R_{I,true}^2 = 1 - \frac{\sum_{i=1}^n [p_i \log p_i + (1-p_i) \log(1-p_i)]}{\sum_{i=1}^n [\bar{p} \log \bar{p} + (1-\bar{p}) \log(1-\bar{p})]} \quad (1.8)$$

เมื่อ p_i คือค่าเริ่มต้นที่ได้จากการจำลอง $\text{logit}(p_i)$ ในข้อ 1.2
 $\bar{p}$ คือค่าเฉลี่ยของ p_i เริ่มต้น $\left(\bar{p} = \frac{p_1 + p_2 + \dots + p_n}{n} \right)$

2. รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error, RMSE) ของค่าประมาณสัมประสิทธิ์การตัดสินใจ ($R_{est,k}^2$) ที่คำนวณจากการทำซ้ำจำนวน 1,000 ครั้ง เมื่อ $k=1, 2, \dots, 1000$ โดยที่

ค่า RMSE คำนวณแยกตามวิธีการที่ใช้ในการคำนวณค่า R^2 และ R_{adj}^2 ดังนี้

$$RMSE_O = \sqrt{\frac{\sum_{k=1}^{1,000} (R_{O,true}^2 - R_{O,est,k}^2)^2}{1,000}}$$

$$RMSE_I = \sqrt{\frac{\sum_{k=1}^{1,000} (R_{I,true}^2 - R_{I,est,k}^2)^2}{1,000}}$$

เมื่อ $RMSE_O$ และ $RMSE_I$ คือค่า RMSE ที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญ และวิธีความควรจะเป็นสูงสุด ตามลำดับ

$R_{O,est,k}^2$ และ $R_{I,est,k}^2$ คือค่าประมาณ R^2 หรือ R_{adj}^2 จากการคำนวณด้วยวิธีกำลังสองน้อยสุด และวิธีความควรจะเป็นสูงสุด ตามลำดับ ที่ได้จากการทำซ้ำครั้งที่ k

3. ร้อยละของค่าประมาณสัมประสิทธิ์การตัดสินใจ (R_O^2 , R_I^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า ($R_{O,adj,MS}^2$, $R_{I,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{I,adj,LM}^2$, $R_{I,adj,SAS_{DEV}}^2$, $R_{I,adj,SAS_{AIC}}^2$) ที่มีค่าอยู่นอกช่วง $[0,1]$ จากการทำซ้ำจำนวน 1,000 ครั้ง

การเปรียบเทียบว่าสัมประสิทธิ์การตัดสินใจแบบใดมีประสิทธิภาพมากที่สุด จะพิจารณาสัมประสิทธิ์การตัดสินใจที่มีค่าความเอนเอียงและความเอนเอียงสัมพัทธ์น้อยที่สุด ร่วมกับการพิจารณาค่า RMSE และร้อยละของสัมประสิทธิ์การตัดสินใจและสัมประสิทธิ์การตัดสินใจที่ปรับค่าที่มีค่าอยู่นอกช่วง $[0,1]$ น้อยที่สุด

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการเลือกใช้สัมประสิทธิ์การตัดสินใจของตัวแบบการถดถอยโลจิสติก ให้เหมาะสมในแต่ละสถานการณ์ของจำนวนตัวแปรอธิบายและขนาดตัวอย่าง
2. เพื่อเป็นแนวทางในการกำหนดขนาดตัวอย่างที่เหมาะสมสำหรับการวิเคราะห์การถดถอยโลจิสติก
3. เพื่อทราบอิทธิพลของการกำหนดตัวแบบการถดถอยโลจิสติกไม่ถูกต้องใน 3 ลักษณะ ที่มีต่อค่าสัมประสิทธิ์การตัดสินใจและสัมประสิทธิ์การตัดสินใจที่ปรับค่า

สัญลักษณ์และคำจำกัดความที่ใช้ในงานวิจัย

ตัวแบบการถดถอยโลจิสติก คือตัวแบบที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตามแบบทวิภาคที่มีค่าเป็นไปได้ 2 ค่า

Odd คืออัตราส่วนของโอกาสที่จะเกิดความสำเร็จต่อโอกาสที่จะเกิดความสำเร็จไม่ได้ในรูป $p/(1-p)$ เมื่อ p คือโอกาสที่จะเกิดความสำเร็จ

Deviance คือตัวสถิติที่ใช้วัดความเหมาะสมของตัวแบบการถดถอย คำนวณจากอัตราส่วนของฟังก์ชันควรจะเป็นของตัวแบบ 2 ตัวแบบ มีการแจกแจงเมื่อใกล้ลัคนันต์แบบไคกำลังสองที่มีองศาเสรีเท่ากับ $n-p$ เมื่อ n คือขนาดตัวอย่างและ p คือจำนวนตัวแปรอธิบายในตัวแบบ

R_O^2 แทนสัมประสิทธิ์การตัดสินใจที่คำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ

- R^2 แทนสัมประสิทธิ์การตัดสินใจที่คำนวณโดยใช้วิธีความควรจะเป็นสูงสุด
- $R_{O,true}^2$ แทนค่าสัมประสิทธิ์การตัดสินใจที่แท้จริงจากการคำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญ
- $R_{I,true}^2$ แทนค่าสัมประสิทธิ์การตัดสินใจที่แท้จริงจากการคำนวณด้วยวิธีความควรจะเป็นสูงสุด
- $R_{O,adj,MS}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยองศาเสรี เมื่อคำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ ที่ศึกษาในงานของมิตเบิร์กและสเชมเปอร์
- $R_{I,adj,MS}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยสถิติดีไวเนียนซ์ เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด ที่ศึกษาในงานของมิตเบิร์กและสเชมเปอร์
- $R_{O,adj,LM}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าของเลียวและแมคกี เมื่อคำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ
- $R_{I,adj,LM}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าของเลียวและแมคกี เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด
- $R_{I,adj,SAS_{DEV}}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยสถิติดีไวเนียนซ์ เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด ที่ใช้ใน PROC LOGISTIC ของโปรแกรม SAS[®]
- $R_{I,adj,SAS_{AIC}}^2$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยหลักเกณฑ์สารสนเทศของ AIC เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด ที่ใช้ใน PROC GENMOD ของโปรแกรม SAS[®]
- $R_{est,k}^2$ ค่าประมาณของสัมประสิทธิ์การตัดสินใจและสัมประสิทธิ์การตัดสินใจที่ปรับค่าสำหรับตัวแบบการถดถอยโลจิสติกที่ได้จากการทำซ้ำครั้งที่ k เมื่อ $k = 1, 2, \dots, 1000$
- $\hat{R}^2$ คือค่ามัธยฐาน (median) ของ $R_{est,k}^2$ จากการทำซ้ำจำนวน 1,000 ครั้ง
- $RMSE_O$ คือรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error) ของ $R_{est,k}^2$ ที่คำนวณด้วยวิธีกำลังสองน้อยสุด จากการทำซ้ำจำนวน 1,000 ครั้ง
- $RMSE_I$ คือรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error) ของ $R_{est,k}^2$ ที่คำนวณด้วยวิธีความควรจะเป็นสูงสุด จากการทำซ้ำจำนวน 1,000 ครั้ง
- out of range(%) หมายถึง ร้อยละของค่าประมาณสัมประสิทธิ์การตัดสินใจ (R_O^2, R_I^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า ($R_{O,adj,MS}^2, R_{I,adj,MS}^2, R_{O,adj,LM}^2, R_{I,adj,LM}^2, R_{I,adj,SAS_{DEV}}^2, R_{I,adj,SAS_{AIC}}^2$) ที่มีค่าอยู่นอกช่วง $[0,1]$ จากการทำซ้ำจำนวน 1,000 ครั้ง

bias หมายถึง ความเอนเอียงของค่าประมาณมัธยฐานที่แตกต่างจากค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง ($\hat{R}^2 - R_{true}^2$)

relative bias หมายถึง ความเอนเอียงสัมพัทธ์ ที่วัดจากค่าสัมบูรณ์ของ
$$\left(\frac{\hat{R}^2 - R_{true}^2}{R_{true}^2} \right) \times 100\%$$

การให้รหัสตัวแบบที่ใช้ในการศึกษา

อักษรย่อ 1 ตัว และตัวเลข 2 ตัว

อักษร M หมายถึงตัวแบบ (Model)

ตัวเลขตัวแรกหลังอักษร M หมายถึง ลักษณะตัวแบบการถดถอย โดยมีความหมายดังต่อไปนี้

เลข 0 หมายถึง ตัวแบบที่กำหนดฟังก์ชันเชื่อมโยงที่แท้จริงเป็นฟังก์ชันโลจิท

เลข 1 หมายถึง ตัวแบบที่กำหนดรูปแบบของตัวแปรอธิบายไม่ถูกต้อง

เลข 2 หมายถึง ตัวแบบที่กำหนดฟังก์ชันเชื่อมโยงเป็นฟังก์ชันคอมพลีเมนต์ลอจิสติก

เลข 3 หมายถึง ตัวแบบที่กำหนดฟังก์ชันเชื่อมโยงเป็นฟังก์ชันโพรบิท

เลข 4 หมายถึง ตัวแบบที่กำหนดให้ตัวแปรอธิบายเชิงปริมาณขาดหายไป

เลข 5 หมายถึง ตัวแบบที่กำหนดให้ตัวแปรอธิบายเชิงคุณภาพขาดหายไป

เลข 6 หมายถึง ตัวแบบที่กำหนดให้ตัวแปรอธิบายเชิงปริมาณและเชิงคุณภาพขาดหายไป

ตัวเลขตัวที่สอง หมายถึง จำนวนตัวแปรอธิบายในตัวแบบการถดถอย โดยมีความหมายดังต่อไปนี้

เลข 1 หมายถึง ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 1 ตัว

เลข 5 หมายถึง ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 5 ตัว

เลข 10 หมายถึง ตัวแบบที่ประกอบด้วยตัวแปรอธิบาย 10 ตัว