หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย

Principle and Using Logistic Regression Analysis for Research

ยุทธ ไกยวรรณ์ [1*]

Yuth Kaiyawan [1*]

บทคัดย่อ

การวิเคราะห์การถดถอยโลจิสติค เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการโลจิสติคที่สร้างขึ้นจากชุดตัวแปรทำนาย ที่เป็นตัวแปรที่มีข้อมูลอยู่ในระดับช่วงเป็นอย่างน้อย โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ และในการวิเคราะห์จะต้องใช้ขนาดตัวแปรทำนายไม่ต่ำกว่า 30 ตัวแปร

การวิเคราะห์การถดถอยโลจิสติคแบ่งออกเป็น 2 ประเภท ได้แก่ (1) การวิเคราะห์การถดถอยโลจิสติคทวิ ใช้กับตัวแปรเกณฑ์ที่แบ่งออกเป็น 2 กลุ่มย่อย เช่น กลุ่มที่ปรากฏเหตุการณ์ที่สนใจ มีค่าเป็น 1 กับกลุ่มที่ไม่ปรากฏเหตุการณ์ที่สนใจ มีค่าเป็น 0 และ (2) การวิเคราะห์การถดถอยโลจิสติคพหุกลุ่ม ใช้กับตัวแปรเกณฑ์ที่มีหลายกลุ่มย่อยหรือมีมากกว่า 2 กลุ่มย่อย เช่น ขนาดสถานศึกษา แบ่งเป็นกลุ่มขนาดใหญ่/กลาง/เล็ก หรือโรงพยาบาลมีมาตรฐานการให้บริการสูง/ปานกลาง/ต่ำ

การวิเคราะห์การถดถอยโลจิสติค ความสัมพันธ์ระหว่างตัวแปรทำนายกับตัวแปรเกณฑ์จึงไม่เป็นความสัมพันธ์เชิงเส้นในการวิเคราะห์จะต้องมีการปรับให้ความสัมพันธ์อยู่ในรูปเชิงเส้น ในรูปของ odds และในการเขียนโมเดลโลจิสติค จะต้องเขียนให้อยู่ในรูป log ของ odds เรียกว่า logit

คำสำคัญ: โลจิสติค, การวิเคราะห์การถดถอย

ABSTRACT

The objective of logistic regression analysis is to predict the occurrence of interested events. By using equation of logistic regression analysis that erected from set of predict or variables. Which variable has data in an interval scale at least. Which the relation between prediction variables must

[1] สาขาการจัดการเทคโนโลยี คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏเพชรบุรี อำเภอเมือง จังหวัดเพชรบุรี 76000

[1] Technology Management, Faculty of Industrial Technology, Phetchaburi Rajabhat University, Muang, Phetchaburi 76000, Thailand.

[*] ผู้นิพนธ์ประสานงาน ไปรษณีย์อิเล็กทรอนิกส์ (corresponding author, e-mail): dr.yuth_go@hotmail.com

be low and analysis must be base on variable size not less than 30 p once p is predictor variable.

Logistics regression analysis was divided into two categories. (1) Binary logistic regression analysis is used for criterion variable that divided into two subgroups such as the group, that showing interested event, will be value 1 with the group, that not showing interested event, will be 0. (2) Multinomial logistic regression is using for criterion variable that divided into several subgroups or more than two subgroups such as size of educational institutes in divided into big group/middle/ small group or hospital may has excellent/medium/minimum services.

In logistic regression analysis, once criterion variable become qualitative variable, has two subgroups or more than two subgroups. The relationship between predictor variable and criterion variable will be nonlinear regression. So that analysis needs to adjust the relationship to be linear regression and logistic model must be writing in form log of odds, that called logit.

Key words: logistic, regression analysis

บทนำ

การวิเคราะห์การถดถอยโลจิสติค (logistic regression analysis) เป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ (qualitative statistical techniques) ที่แตกต่างไปจากเทคนิคการวิเคราะห์เชิงปริมาณ (quantitative techniques) อย่างน้อย ก็เรื่องของข้อมูลที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ซึ่งก็คือ เป็นตัวแปรเชิงกลุ่มนั่นเอง การวิเคราะห์การถดถอยโลจิสิคแบ่งเป็น 2 ประเภท คือ (1) การวิเคราะห์การถดถอยโลจิสติคทวิ (binary logistic regression analysis) และ (2) การวิเคราะห์การถดถอยโลจิสติคพหุกลุ่ม (multinomial logistic regression analysis) การวิเคราะห์การถดถอยโลจิสติคทั้ง 2 ประเภท แตกต่างกันในด้านตัวแปรตาม โดยที่การวิเคราะห์การถดถอยโลจิสติคทวิใช้กับตัวแปรตามที่แบ่งออกเป็น 2 กลุ่มย่อย (dichotomous variable) มี 2 ค่า คือมีค่าเป็น 0 กับ 1 เช่น กลุ่มที่มีเหตุการณ์กับกลุ่มที่ไม่มีเหตุการณ์ ส่วนการวิเคราะห์โลจิสติคแบบพหุกลุ่มใช้กับตัวแปรตามที่มีหลายค่ามากกว่า 2 กลุ่ม (polytomous variable) เช่น โรงพยาบาลมีมาตรฐานการให้บริการ สูง ปานกลาง และต่ำ

การวิเคราะห์โลจิสติคมีเป้าหมายก็คือ เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งก็คือตัวแปรเกณฑ์ โดยอาศัยสมการโลจิสติคที่สร้างขึ้น จากชุดตัวแปรทำนาย (x's) ที่มีข้อมูลเป็นตัวแปรที่มีข้อมูลอยู่ในระดับช่วง (interval scale) เป็นอย่างน้อย หากเป็นข้อมูลเชิงกลุ่มจะต้องแปลงเป็นตัวแปรทวิ ที่มีค่า 0 กับ 1 ก่อน โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ โดยใช้เกณฑ์ค่า r ไม่เกิน .65 ถ้าใช้เกณฑ์ของ Burns and Grove (1993) หรือถ้าใช้เกณฑ์ของ Stevens (1996) ค่า r ไม่เกิน .80 ซึ่งถ้าหากเกิดความสัมพันธ์กันสูงจะทำให้เกิดปัญหา multicolinearity และในการวิเคราะห์จะต้องใช้ขนาดตัวอย่างหรือ n มากกว่าหรือเท่ากับ 30 เท่าของจำนวนตัวแปรทำนาย (กัลยา, 2549)

วัตถุประสงค์การวิเคราะห์การถดถอยโลจิสติค

การวิเคราะห์การถดถอยโลจิสติค มีวัตถุประสงค์เพื่อศึกษาว่า ตัวแปรอิสระหรือตัวแปรทำนายใดบ้างที่สามารถอธิบายตัวแปรเกณฑ์ (ตัวแปรตาม) ซึ่งเป็นตัวแปรทวิหรือตัวแปรพหุกลุ่ม โดยอาจจะมีประเด็นปัญหาของการศึกษาดังนี้ (ศิริชัย, 2550)

1. ตัวแปรอิสระใดบ้างที่สามารถใช้อธิบายโอกาสการเกิดเหตุการณ์หรือการไม่เกิดเหตุการณ์ที่สนใจตามตัวแปรตามหรือตัวแปรเกณฑ์ พร้อมทั้งศึกษาระดับความสัมพันธ์ของตัวแปรทำนายแต่ละตัว

2. เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ จากสมการโลจิสติคที่เหมาะสม โดยเลือกตัวแปรที่เหมาะสมเพื่อทำให้เปอร์เซ็นต์ของความถูกต้องในการทำนายมีค่าสูงสุด

ข้อตกลงเบื้องต้นการวิเคราะห์การถดถอยโลจิสติค

การวิเคราะห์การถดถอยโลจิสติค มีข้อตกลงเบื้องต้น ดังนี้

1. ตัวแปรอิสระหรือตัวแปรทำนาย (x's) เป็นตัวแปรที่ระดับข้อมูลอยู่ในระดับช่วง (interval scale) เป็นอย่างต่ำ กรณีที่เป็นข้อมูลเชิงกลุ่มให้แปลงเป็นตัวแปรหุ่น (dichotomous variable) ที่มีค่าเป็น 0 กับ 1 เท่านั้น ส่วนตัวแปรเกณฑ์หรือตัวแปรตาม กรณีที่เป็นการวิเคราะห์โลจิสติคแบบทวิ (binary logistic regression) จะกำหนด 2 ค่า คือ 0 กับ 1 ส่วนกรณีการวิเคราะห์ โลจิสติคพหุกลุ่ม (multinomial logistic regression) จะกำหนดตามจำนวนกลุ่มของตัวแปรเกณฑ์

2. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์หรือไม่มีความสัมพันธ์กัน นั่นคือ $e = 0$ (เพชรน้อย, 2549)

3. ตัวแปรอิสระไม่มีความสัมพันธ์กันหรือไม่เกิดปัญหา multicollinearity (กัลยา, 2549) ทั้งนี้จะใช้เกณฑ์ความสัมพันธ์เหมือนกับการวิเคราะห์การถดถอยพหุ โดยถ้าใช้เกณฑ์ของ Burns and Grove (1993) จะใช้ค่า r ไม่เกิน .65 และถ้าใช้เกณฑ์ของ Stevens (1996) ใช้ค่า r ไม่เกิน .80

4. การวิเคราะห์การถดถอยโลจิสติคจะต้องใช้ขนาดตัวอย่าง n มากกว่าการวิเคราะห์การถดถอยแบบปกติ โดยจะใช้ขนาดตัวอย่างเท่ากับ $n \geq 30\ p$ โดยที่ p คือ จำนวนตัวแปรทำนาย (กัลยา, 2549)

โมเดลการวิเคราะห์การถดถอยโลจิสติค

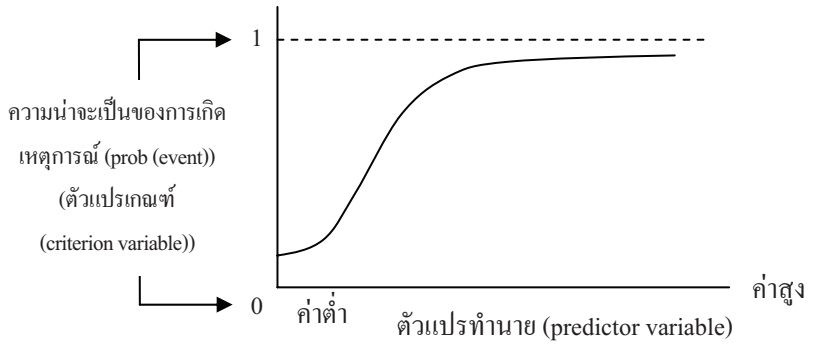
1. โมเดลการวิเคราะห์การถดถอยโลจิสติคทวิ (binary logistic regression analysis model)

1.1 กรณีตัวแปรทำนาย 1 ตัว

ในการวิเคราะห์การถดถอยอย่างง่าย (simple regression analysis) สมการที่แสดงความสัมพันธ์ระหว่าง x กับ y จะอยู่ในรูปเชิงเส้น ดังนี้

$$y = b_0 + b_1x + e$$

แต่สำหรับการวิเคราะห์โลจิสติคที่เป็นแบบทวิ ตัวแปรตามหรือตัวแปรเกณฑ์ (y) มี 2 ค่า คือ ไม่เกิดเหตุการณ์ (y = 0) หรือเกิดเหตุการณ์ (y = 1) มีความสัมพันธ์กับตัวแปรทำนาย (x) ไม่อยู่ในรูปเชิงเส้น ทั้งนี้เพราะตัวแปรตามมี 2 ค่า คือ 0 กับ 1 จึงเป็นไปไม่ได้ที่ความสัมพันธ์จะอยู่ในรูปเส้นตรง ซึ่งความสัมพันธ์ของตัวแปรของการวิเคราะห์โลจิสติคจะอยู่ในรูปคล้ายตัว s ดังภาพต่อไปนี้



ภาพที่ 1 ฟังก์ชั่นโลจิสติค (logistic function)

โดยที่ $p(y) = \frac{1}{1+e^{-f(x)}}$

หรือ $\frac{1}{1+e^{-(b_0+b_1x)}}$ หรือ $\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$

เมื่อ p(y) = ความน่าจะเป็นของการเกิดเหตุการณ์ y

e = exponential function (e = 2.71828)

f(x) = ฟังก์ชั่นของตัวแปรทำนาย

สมมติให้

P_y = ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (y = 1)

Q_y = ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ (y = 0)

z = linear combination ของตัวแปรทำนาย (กรณีตัวแปรทำนาย 1 ตัวแปร)

$z = b_0 + b_1x$

จะได้ $p(y) = \frac{1}{1+e^{-z}}$

$= \frac{e^z}{1+e^z}$

$= \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$

และ $Q_y = 1 - P_y$

หรือ $Q_y = 1-\frac{e^z}{1+e^z} = \frac{1+e^z-e^z}{1+e^z} = \frac{1}{1+e^z}$

1.2 กรณีตัวแปรทำนายมากกว่า 1 ตัว (ตัวแปรทำนาย > 1 ตัว)

ในการวิเคราะห์เมื่อตัวแปรทำนายมีมากกว่า 1 ตัว จะได้ฟังก์ชั่นดังนี้

$$P_y = \frac{e^{b_0+b_1x_1+...+b_px_p}}{1+e^{b_0+b_1x_1+...+b_px_p}}$$

เมื่อ P_y = ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

จะได้ Qy หรือความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ ดังนี้

$Q_y = 1 - Py$

หรือ $Q_y = 1-\left(\frac{e^{b_0+b_1x_1+...+b_px_p}}{1+e^{b_0+b_1x_1+...+b_px_p}}\right)$

จากความสัมพันธ์ระหว่างตัวแปรทำนายกับตัวแปรเกณฑ์ของการวิเคราะห์ถดถอยโลจิสติคไม่เป็นรูปเชิงเส้น จึงต้องมีการปรับให้ความสัมพันธ์ให้อยู่ในรูปเชิงเส้น ในรูปแบบของ odds หรือ odd ratio

odds หรือ odd ratio หมายถึง อัตราส่วน

ระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจ (y = 1) กับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ (y = 0) หรือจะได้

$$\text{odds} = \frac{P_y}{Q_y} \quad \begin{matrix}\text{(โอกาสเกิดเหตุการณ์ที่สนใจ)} \\ \text{(โอกาสไม่เกิดเหตุการณ์ที่สนใจ)}\end{matrix}$$

ค่าของ odds จะแสดงถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจ เป็นกี่เท่าของโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ เช่น odds ของการโยนเหรียญ 1 ครั้งเท่ากับ $\frac{0.5}{0.5} = 1$ หรือ ถ้า odds มีค่าเท่ากับ 2.5 แสดงว่า โอกาสที่จะเกิดเหตุการณ์ที่สนใจเป็น 2.5 เท่าของโอกาสที่จะไม่เกิด ถ้า odds มีค่าเท่ากับ 1 โอกาสที่จะเกิดเหตุการณ์ที่สนใจกับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจเท่ากัน นั่นคือ ถ้า odds มีค่ามากกว่า 1 แสดงว่า โอกาสที่จะเกิดเหตุการณ์ที่สนใจนั้นมากกว่าโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ

การเขียนโมเดลโลจิสติค จะอยู่ในรูป log ของ odds เรียกว่า logit หรือ logistic response function ซึ่ง logit เขียนในรูปสมการดังนี้

เมื่อ $\text{odds} = \frac{P_y}{Q_y}$

จะได้ log ของ odds หรือจะเรียก log ของ odds ว่า logit ดังนี้

log(odds)

ดังนั้นจะได้ $\log\left(\frac{P_y}{Q_y}\right)$

เมื่อ $Q_y = 1 - P_y$ จะได้ $\log\left(\frac{P_y}{1-P_y}\right) = b_0 + b_1x_1 + \ldots + b_px_p$

หรือ log(odds) หรือ logit $= b_0 + b_1x_1 + \ldots + b_px_p$

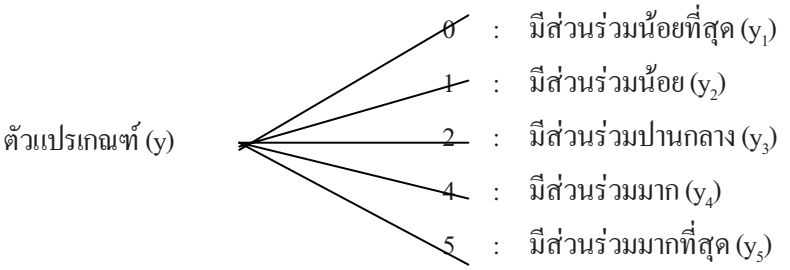
เมื่อได้ log ของ odds ratio หรือ logit แล้ว รูปแบบของตัวแปรเกณฑ์ จึงสามารถทำนายได้ด้วยชุดของตัวแปรทำนายเชิงเส้นตรง

สำหรับการทำนายค่า y ที่เป็น P_y ในการวิเคราะห์ถดถอยโลจิสติคจะใช้สมการ

$$Py = \frac{e^{b_0+b_1x_1+\ldots+b_px_p}}{1+e^{b_0+b_1x_1+\ldots+b_px_p}}$$

ตามวิธี maximum likelihood (ศิริชัย, 2549) ในขณะที่การทำนายค่า y ในการวิเคราะห์การถดถอยปกติจะใช้วิธี least square จากสมการ $y = b_0 + b_1x_1 + \ldots + b_px_p$ (กัลยา, 2549)

2. โมเดลการวิเคราะห์การถดถอยโลจิสติคพหุกลุ่ม (multinomial logistic regression model) ในกรณีที่ตัวแปรเกณฑ์ (y) เป็นตัวแปรเชิงกลุ่มมีค่ามากกว่า 2 ค่า การวิเคราะห์การถดถอยโลจิสติคจะใช้เทคนิค multinomial logistic regression เช่น



ภาพที่ 2 แสดงตัวแปรเกณฑ์ (y) กรณีมีมากกว่า 2 กลุ่ม

ถ้าตัวแปรเกณฑ์ (y) หมายถึง ระดับการมีส่วนร่วมในการพัฒนาองค์กร โดยแบ่งออกเป็น 5 ระดับ คือ

ตัวแปรทำนายอาจเป็นดังนี้ การศึกษา อาชีพ อายุ ตำแหน่งทางสังคม ถ้าผู้วิจัยสร้างโมเดลแสดงความสัมพันธ์ระหว่างตัวแปรเกณฑ์ (y) กับตัวแปรทำนาย (x's) จะทำให้ทราบว่าตัวแปรทำนายใดบ้างที่มีความสัมพันธ์หรือมีอิทธิพลต่อกลุ่มตัวแปรเกณฑ์ที่กำหนด (กัลยา, 2549)

ลักษณะของความสัมพันธ์ระหว่างตัวแปรทำนายกับตัวแปรเกณฑ์ของการวิเคราะห์ โลจิสติคพหุกลุ่ม ดังนี้

ในกรณีตัวแปรเกณฑ์ (y) มีค่า 2 ค่า จะเป็น binary logistic model ดังนี้

$$\log\left(\frac{P_y}{Q_y}\right) \text{ เมื่อ } Q_y = 1\text{-}P_y$$

$$\text{จะได้ } \log\left(\frac{P_y}{1-P_y}\right) = b_0 + b_1x_1 + \ldots + b_px_p$$

แต่เมื่อตัวแปรเกณฑ์มีมากกว่า 2 ค่า เช่น K > 2 จะได้ logit จำนวนเท่ากับ (K-1) และจะนำ logit แต่ละค่าเปรียบเทียบกับกลุ่มที่เป็นฐาน (baseline category) ซึ่งค่าสัมประสิทธิ์ทั้งหลายของตัวแปรสำหรับกลุ่มที่เป็นฐานจะเท่ากับ 0 เพื่อเป็นฐานในการเปรียบเทียบกับค่าของกลุ่มอื่น และหากกรณีที่ตัวแปรเกณฑ์มี 3 หรือ 4 โดยที่กลุ่มที่เป็นฐานคือ K และเมื่อเมื่อเปรียบเทียบกับกลุ่มที่ i จะได้ logit model ดังนี้

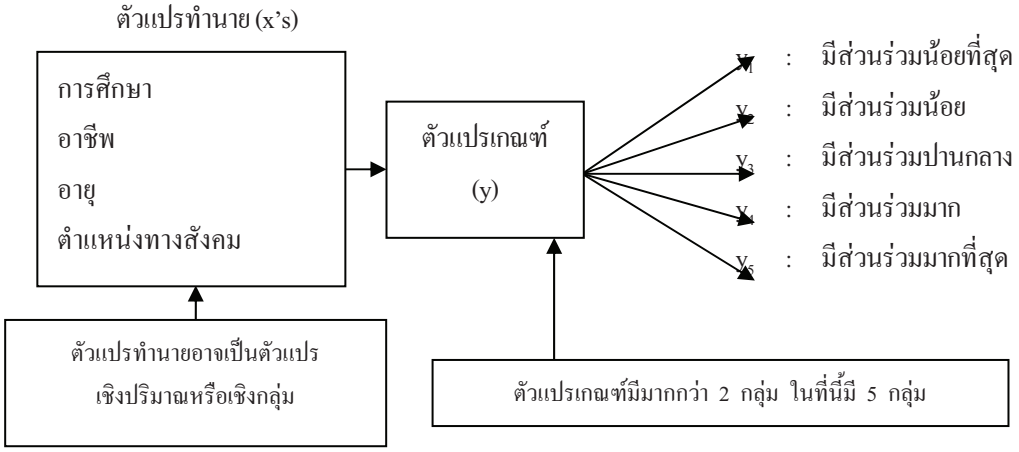
$$\log\left(\frac{P(\text{กลุ่ม i})}{P(\text{กลุ่ม K})}\right) = b_{io} + b^{i1}x_1 + \ldots + b_{ip}x_p$$

สัมประสิทธิ์ b_{io}, b_{i1}, b_{i2}, … , bip ของกลุ่มที่ i และกลุ่มที่เป็นฐาน (baseline category) จะมีค่าเป็น

$b_0 = b_1 = \ldots = b_p = 0$ การวิเคราะห์จะให้ผล ดังนี้

ถ้าตัวแปรเกณฑ์ (y) มี 3 ค่า หรือ K = 3 จะได้ผลลัพธ์สัมประสิทธิ์ 2 เซ็ทหรือชุด มาจาก K-1 แต่ baseline category จะมี 3 ค่าตามตัวแปรเกณฑ์ (y) ใน 2 เซ็ท

ชุดที่ 1 แสดงค่าสัมประสิทธิ์ของ y = 1 เปรียบเทียบกับ y = 3



ภาพที่ 3 แสดงความสัมพันธ์ของตัวแปรทำนายกับตัวแปรเกณฑ์การวิเคราะห์การถดถอยโลจิสติคพหุกลุ่ม

และชุดที่ 2 แสดงค่าสัมประสิทธิ์ของ y = 2 เปรียบเทียบกับ y = 3

ตัวอย่างการเปรียบเทียบกลุ่มหรือของสมมติฐานผู้วิจัยสนใจ นักเรียนมัธยมศึกษาปีที่ 3 จะเลือกเข้ามัธยมศึกษาปีที่ 4 โรงเรียนใดใน 3 โรงเรียน ได้แก่ Aschool Bschool และ Cschool ในการเลือกนี้ผู้วิจัยสงสัยว่า เพศ (sex) ของนักเรียนจะมีผลต่อการเลือกโรงเรียนหรือไม่ โดยให้เพศหญิงเป็นกลุ่มที่เป็นฐาน (baseline category) ค่าสัมประสิทธิ์ของเพศหญิงจะเท่ากับ 0 ผู้วิจัยจะได้ logit 2 ค่าที่ไม่ซ้ำซ้อนกันดังนี้

$$g_1 = \left(\frac{P(\text{Aschool})}{P(\text{Cschool})}\right) = b_{10} + b_{11}\ (\text{ชาย})$$

เลขกำกับที่ของตัวแปร

$$g_2 = \left(\frac{P(\text{Bschool})}{P(\text{Cschool})}\right) = b_{20} + b_{21}\ (\text{ชาย})$$

เลขกำกับที่ของ logit

การตรวจสอบความเหมาะสมของสมการถดถอยโลจิสติค

การตรวจสอบความเหมาะสมของสมการถดถอยโลจิสติค มีการตรวจสอบหลายวิธี ดังนี้

1. พิจารณาค่าความเป็นไปได้ (likelihood value)

พิจารณาค่าความเป็นไปได้ เพื่อวัดค่าความเหมาะสมของสมการโลจิสติค จะศึกษาจากค่า -2LL (-2 log likelihood) ซึ่งเป็นค่ามาจาก log likelihood ที่คูณด้วย -2 เพื่อต้องการให้ค่าที่ได้มีการแจกแจงมีลักษณะเป็นการแจกแจง χ^2 สำหรับการทดสอบนัยสำคัญทางสถิติ

การพิจารณาค่า -2LL ถ้ามีค่าต่ำ สมการโลจิสติค มีความเหมาะสมที่สุดในการทดสอบนัยสำคัญความเหมาะสมของสมการโลจิสติค ใช้สถิติ χ^2-test

การทดสอบ model Chi-square ที่ df = p (จำนวนตัวแปรทำนาย) เป็นการทดสอบสมมติฐานดังนี้

H_0 : สัมประสิทธิ์ถดถอยโลจิสติคทุกตัวมีค่า = 0

หรือ $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : มี $\beta_i \neq 0$ อย่างน้อย 1 ค่า; i = 1, 2,…, p

การทดสอบถ้า χ^2 มีนัยสำคัญทางสถิติหรือยอมรับ H_1 แสดงว่า ชุดตัวแปรทำนาย (x's) สามารถร่วมกันทำนายโอกาสของการเกิดเหตุการณ์ที่สนใจ (y = 1) ได้ด้วยความเชื่อ (1 - α) x 100%

2. พิจารณาสถิติทดสอบความเหมาะสมของ Hosmer and Lemeshow

จะใช้ทดสอบความเหมาะสม model ดังนี้

$$p(y) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + \ldots + b_px_p)}}$$

สมมติฐานที่ทดสอบ คือ

H_0 : model เหมาะสม

H_1 : model ไม่เหมาะสม

ในการทดสอบหาก χ^2 ไม่มีนัยสำคัญทางสถิติหรือยอมรับ H_0 แสดงว่า model มีความเหมาะสม

การทดสอบนัยสำคัญของสัมประสิทธิ์การถดถอยโลจิสติคของตัวแปรทำนายแต่ละตัว

การทดสอบนัยสำคัญของสัมประสิทธิ์ถดถอยโลจิสติค ทดสอบด้วยสถิติทดสอบ 2 ตัว ได้แก่

1. สถิติทดสอบของวอลด์ (Wald statistic)

วอลด์ (Wald statistic) เป็นการทดสอบสมมติฐานที่กำหนด ดังนี้

H_0 : ตัวแปรทำนาย (x's) ไม่มีผลต่อการเปลี่ยนแปลงของ odds ratio

หรือ $H_0 : \beta_i = 0; i = 1, 2, \ldots, p$

H_1 : มี $\beta_i \neq 0$

ในการทดสอบถ้าผลการทดสอบยอมรับ H_0 แสดงว่า ตัวแปรทำนาย i ไม่มีผลต่อการเปลี่ยนแปลง odds ratio ดังนั้นจึงไม่มีผลต่อความน่าจะเป็นของการเกิดเหตุการณ์นั้น และถ้าทดสอบพบว่า มีนัยสำคัญทางสถิติหรือยอมรับ H_1 และค่าสัมประสิทธิ์เป็นบวก (+) แสดงว่าตัวแปรทำนายนั้นมีผลต่อการเพิ่มความน่าจะเป็นของการเกิดเหตุการณ์ และถ้าหากค่าสัมประสิทธิ์เป็นลบ (-) แสดงว่าตัวแปรทำนายนั้นลดความน่าจะเป็นของการเกิดเหตุการณ์ (ศิริชัย, 2549)

สถิติทดสอบของวอลด์ (Wald test) จะมีการแจกแจงแบบ χ^2 และ df = 1

สถิติทดสอบ คือ Wald หรือ $w = \left[\frac{b_o}{SE(b_o)}\right]^2$

หรือทดสอบฟังก์ชั่น

$w = \text{constant} + b_1x_1 + b_2x_2 + \ldots + b_px_p$

หรือ $w = b_o + b_1x_1 + b_2x_2 + \ldots + b_px_p$

ผลจากการวิเคราะห์ด้วยโปรแกรมสำเร็จรูปทางสถิติค่าสัมประสิทธิ์ความถดถอยโลจิสติคของตัวแปรทำนายแต่ละตัว(β_i) และสถิติ wald

2. สถิติทดสอบความเหมาะสม สัมประสิทธิ์ของ model

การทดสอบความเหมาะสม สัมประสิทธิ์ของ model เป็นการทดสอบค่าสัมประสิทธิ์ความถดถอยโลจิสติคในรูปของค่าสถิติ χ^2 ซึ่งมีด้วยกัน 3 ค่า ได้แก่

ค่า step, block และ model และจากผลการวิเคราะห์ ถ้าพบว่าค่า χ^2 ใน step block และ model มีค่า χ^2 เท่ากันทั้ง 3 ค่า และมีนัยสำคัญทางสถิติแสดงว่า ตัวแปรทำนายที่เพิ่มเข้าไปใน model นั้นมีความเหมาะสมดี

สมมติฐานที่กำหนดเป็นการทดสอบความเหมาะสมของสัมประสิทธิ์นี้คือ

H_0 : model ไม่ขึ้นอยู่กับตัวแปรทำนายทั้ง p ตัว ($x_1, x_2, \ldots, x_p$)

หรือ $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p$

H_1 : model ขึ้นอยู่กับตัวแปรทำนายนี้อย่างน้อย 1 ตัว

ผลการทดสอบ ถ้า model χ^2 มีนัยสำคัญจะยอมรับ H1

3. สถิติทดสอบระดับความสัมพันธ์

สถิติทดสอบระดับความสัมพันธ์ ดังนี้

3.1 สถิติทดสอบ Cox & Snell R square หรือ R^2_{cs}

สถิติทดสอบ Cox & Snell R square เป็นการพิจารณาหรือตรวจสอบความสอดคล้องของ model หรือเปอร์เซ็นต์ที่สามารถอธิบายความแปรปรวนหรือความผันแปรในการวิเคราะห์ถดถอยโลจิสติค ปกติค่า Cox & Snell R square หรือ R^2 มีค่าน้อยกว่า 1 (< 1) เสมอ (กัลยา, 2551) ถ้าคิดเป็นเปอร์เซ็นต์ให้คูณด้วย 100 สถิติค่านี้จะคล้ายกับค่า R^2 ในการวิเคราะห์ถดถอยพหุปกติ

3.2 สถิติทดสอบ Nagelkerke R square หรือ Nagelkerke R^2 (R^2_N) สถิติค่า R^2_N จะมีลักษณะเหมือนกับ R^2_{cs} แต่จะมีค่ามากกว่า R^2_{cs} เสมอ ค่า R^2_N คิดเป็นเปอร์เซ็นต์ให้คูณด้วย 100

ค่า R^2 ของ Cox & Shell และ Nagelkerke เป็นค่า R^2 เทียม (Pseudo R^2) ซึ่งเป็นค่าเปอร์เซ็นต์

ที่สามารถอธิบายความผันแปรในการวิเคราะห์การถดถอยโลจิสติค

วิธีเลือกตัวแปรทำนายเข้าวิเคราะห์

ในการวิเคราะห์ความถดถอยโลจิสติคเป็นการวิเคราะห์เพื่อทำนายโอกาสที่เหตุการณ์ที่สนใจ (y = 1) จะเกิดขึ้น และสมการถดถอยโลจิสติคที่ดี จะต้องประกอบด้วยตัวแปรทำนายที่เหมาะสมที่จะทำให้ค่าทำนายโอกาสที่จะเกิดนั้นใกล้เคียงกับความเป็นจริง ในการเลือกตัวแปรทำนายเข้าวิเคราะห์ เพื่อให้ได้สมการโลจิสติคที่ดีนั้น มีวิธีเลือก 3 วิธี ซึ่งก็คล้ายกับการวิเคราะห์ถดถอยเชิงพหุปกติ ดังนี้

1) enter method

2) forward method มีวิธีย่อย 3 วิธีคือ

2.1) วิธี forward stepwise : likelihood ratio

2.2) วิธี forward stepwise : wald

2.3) วิธี forward stepwise : condition

3) backward method มีวิธีย่อย 3 วิธี คือ

3.1) วิธี backward stepwise : likelihood ratio

3.2) วิธี backward stepwise : wald

3.3) วิธี backward stepwise : condition

รายละเอียดแต่ละวิธีดังนี้

(1) enter method

วิธี enter methhod เป็นวิธีที่เลือกตัวแปรทำนายทั้งหมด เข้าสมการถดถอยโลจิสติคพร้อมกันในขั้นตอนเดียว ในการพิจารณาตัวแปรทำนายที่เหมาะสมที่จะนำมาใช้ทำนายใน model ผู้วิจัยจะต้องเป็นผู้ตัดสินใจเองว่าตัวแปรทำนายตัวใดบ้างที่มีความสัมพันธ์กับตัวแปรเกณฑ์ หรือควรจะอยู่ในสมการความถดถอยโลจิสติค โดยพิจารณาจากค่าสถิติทดสอบ ถ้ามีนัยสำคัญทางสถิติ ก็ถือว่าตัวแปร ทำนายนั้นควรจะอยู่ในสมการความถดถอยโลจิสติค

(2) forward method

วิธี forward method เป็นการวิเคราะห์ถดถอยโลจิสติคแบบเดินหน้า วิธีการนี้ จะคัดเลือกตัวแปรทำนายที่อธิบายความผันแปรของตัวแปรเกณฑ์ได้สูงสุด และมีนัยสำคัญทางสถิติเข้าสมการก่อน จากนั้นจึงเลือกตัวแปรทำนายที่อธิบายความผันแปรของตัวแปรเกณฑ์ได้อันดับรองลงมา และมีนัยสำคัญทางสถิติเข้าสมการ ตามลำดับ การนำตัวแปรทำนายเข้าสมการจะทำเช่นนี้เรื่อย ๆ ไป จนกระทั่งไม่มีตัวแปรทำนายใดที่อธิบายความผันแปรของตัวแปรเกณฑ์อย่างมีนัยสำคัญทางสถิติอีกแล้ว

การนำตัวแปรทำนายเข้าสมการแบบ forward method มีวิธีย่อย ๆ อีก 3 วิธี ได้แก่

(2.1) วิธี forward stepwise: likelihood ratio

วิธีนี้บางทีเรียกว่า forward LR วิธีนี้จะเริ่มจากการนำตัวแปรทำนายเข้าสมการทีละ 1 ตัว โดยที่ตัวแปรทำนายที่เลือกเข้าสมการทำให้ค่าทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจถูกต้องมากขึ้น เกณฑ์ในการพิจารณาเลือกตัวแปรทำนายเข้าสมการคือ ค่าแสดงความสัมพันธ์ที่มากที่สุดก่อน และมีนัยสำคัญทางสถิติ เมื่อนำตัวแปรทำนายเข้าสมการแล้วจะมีการตรวจสอบตัวแปรทำนายนั้นว่า ควรจะถูกตัดออกหรือควรจะคงอยู่ในสมการ โดยพิจารณาจากอัตราส่วนความเป็นไปได้หรือการเปลี่ยนแปลงของ -2LL (-2 likelihood ratio) ถ้าค่า -2LL ลดลงแสดงว่าตัวแปรทำนาย ควรจะคงอยู่ในสมการ

(2.2) วิธี forward stepwise: wald

วิธีนี้เหมือนกับวิธี forward LR ทุกประการ

เพียงแต่จะพิจารณาจากค่าสถิติของ wald (wald statistic) เท่านั้น

(2.3) วิธี forward stepwise: condition

วิธีนี้จะเหมือนกับวิธี forward LR แตกต่างกันตรงที่ วิธี forward LR เป็นวิธีที่ไม่มีเงื่อนไข (unconditional) ส่วนวิธีนี้จะมีเงื่อนไข (condition) ความแตกต่างของแบบมีเงื่อนไขและไม่มีเงื่อนไขมีดังนี้

1) แบบมีเงื่อนไข ให้ใช้กับตัวอย่างขนาดตัวอย่างเล็ก แบบไม่มีเงื่อนไขใช้กับตัวอย่างขนาดใดก็ได้

2) แบบไม่มีเงื่อนไข มีการควบคุมปัจจัยอื่น ๆ ที่คาดว่าจะมีอิทธิพลต่อโอกาสที่จะเกิดเหตุการณ์ที่สนใจ เช่น ถ้าผู้วิจัยคาดว่าการดื่มแอลกอฮอล์และจำนวนปีที่ดื่มแอลกอฮอล์ทำให้คนเป็นโรคตับแข็ง ตัวแปรเกณฑ์ (y) คือ

$$y = \begin{cases} 1 : \text{เป็นโรคตับแข็ง} \\ 0 : \text{ไม่เป็นโรคตับแข็ง} \end{cases}$$

ส่วนตัวแปรทำนายคือ x_1 = ดื่มแอลกอฮอล์ และ x_2 = จำนวนปีที่ดื่มแอลกอฮอล์

$$x_1 = \begin{cases} 1 : \text{ดื่มแอลกอฮอล์} \\ 0 : \text{ไม่ดื่มแอลกอฮอล์} \end{cases}$$

x_2 = จำนวนปีที่ดื่มแอลกอฮอล์

แบบไม่มีเงื่อนไขจะเลือกตัวอย่างที่เป็นคนไข้ที่เป็นโรคตับแข็งและไม่เป็นโรคตับแข็งมาแล้วศึกษาการดื่มแอลกอฮอล์และจำนวนปีที่ดื่มโดยไม่มีการควบคุมปัจจัยอื่น ๆ ที่คาดว่าจะส่งผลต่อโอกาสการเกิดเหตุการณ์ที่สนใจ เช่น กรรมพันธุ์ โรคพื้นฐานที่ส่งผลต่อการเกิดโรคตับแข็ง เช่น ไวรัสตับบี (B) ปริมาณการดื่ม ความแรงของดีกรี แต่ถ้าเป็นแบบมีเงื่อนไขจะต้องมีการควบคุมปัจจัยดังกล่าว

(3) backward method

วิธี backward method เป็นวิธีที่นำตัวแปรทำนายทั้ง p ตัว ($x_1, x_2, x_3, \ldots, x_p$) เข้าสมการพร้อมกันก่อนจากนั้นพิจารณาตัวแปรทำนายที่อธิบายความผันแปรของตัวแปรเกณฑ์ได้น้อยที่สุดออกจากสมการก่อน ทำเช่นนี้ไปเรื่อยๆ จนกระทั่งเหลือตัวแปรทำนายที่สามารถอธิบายความผันแปรของตัวแปรเกณฑ์ได้อย่างมีนัยสำคัญทางสถิต การนำตัวแปรทำนายเข้าสมการวิธีนี้มีวิธีย่อยอีก 3 วิธี ได้แก่

(3.1) วิธี backward stepwise: likelihood ratio

วิธีนี้บางทีเรียกว่า backward LR ซึ่งเป็นวิธีตรงข้ามกับวิธี forward stepwise: likelihood ratio ซึ่งเป็นวิธีที่นำตัวแปรทำนายทั้งหมด p ตัว ($x_1, x_2, x_3, \ldots, x_p$) เข้าสมการ แล้วพิจารณาว่าจะนำตัวแปรทำนายตัวใดออกจากสมการ โดยพิจารณานำออกทีละ 1 ตัว โดยพิจารณาจากเกณฑ์การนำตัวแปรทำนายออกจากสมการคือ จะนำตัวแปรทำนายที่ไม่มีผลต่อการทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ตัวแปรทำนายตัวแรกที่จะนำออกจากสมการจะเป็นตัวแปรที่ไม่มีผลต่อการทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจน้อยที่สุด สำหรับเกณฑ์การนำตัวแปรทำนายออกจากสมการจะพิจารณาจกอัตราส่วนความเป็นไปได้หรือจากการเปลี่ยนแปลงของ -2LL เหมือนวิธี forward stepwise: likelihood ratio

(3.2) วิธี backward stepwise: wald

วิธีนี้จะเหมือน backward LR ทุกประการ เพียงแต่จะพิจารณาจากค่าสถิติของ wald (wald statistic) เท่านั้น

(3.3) วิธี backward stepwise: condition

วิธีนี้จะเหมือน backward LR ทุกประการ

แตกต่างกันตรงที่วิธี backward LR เป็นวิธีที่ไม่มีเงื่อนไข (unconditional) ส่วนวิธีนี้จะมีเงื่อนไข

ขั้นตอนการวิเคราะห์การถดถอยโลจิสติค

ขั้นตอนที่ 1 สร้างกรอบแนวคิดการวิจัย และกำหนดระดับการวัดของตัวแปรทำนายและตัวแปรเกณฑ์

ขั้นตอนที่ 2 เลือกตัวแปรทำนายที่คาดว่าจะส่งผลต่อตัวแปรเกณฑ์ทั้งนี้การเลือกตัวแปรทำนายสามารถเลือกได้ทีละ 1 ตัว หรือมากกว่าก็ได้

ขั้นตอนที่ 3 ตรวจสอบค่าผิดปกติของตัวแปรทำนายแต่ละตัวทุกตัว ($x_1, x_2,\ldots, x_p$)

ขั้นตอนที่ 4 สร้างสมการ ดังนี้

$$Py = \frac{e^{b_0+b_1x_1+\ldots+b_px_p}}{1+e^{b_0+b_1x_1+\ldots+b_px_p}}$$

ตรวจสอบความถูกต้องเหมาะสมของสมการของ Hosmer and Lemeshow โดยพิจารณาค่าสถิติ χ^2 ถ้าไม่มีนัยสำคัญทางสถิติหรือยอมรับ H_0 แสดงว่า model มีความเหมาะสมดี นอกจากนี้ยังต้องพิจารณาจากค่า pseudo R^2 หรือ R^2 เทียมของ Cox & Snell และ Nagelkerke ซึ่งในการวิเคราะห์การถดถอยแบบปกติจะพิจารณาจากค่า R^2 แต่การวิเคราะห์ถดถอยโลจิสติคจะพิจาณาจากค่า R^2 เทียม (pseudo R^2)

ขั้นตอนที่ 5 ตรวจสอบเงื่อนไขการวิเคราะห์ความถดถอยโลจิสติค

ขั้นตอนที่ 6 วิเคราะห์ข้อมูล

1) เพื่อสร้างสมการถดถอยโลจิสติค

2) ถ้ามีวัตถุประสงค์เพื่อทำนาย case ใหม่ จะใช้สมการ ดังนี้

P(y = 1) หรือ P(y = เกิดเหตุการณ์ที่สนใจ)

$$= \frac{e^{b_0+b_1x_1+\ldots+b_px_p}}{1+e^{b_0+b_1x_1+\ldots+b_px_p}}$$

การทำนายความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจของ case ใหม่ เมื่อทราบค่าตัวแปรทำนาย

ถ้า P (ของ case ใหม่) > 0.5 จะให้เป็น y = 1 หรือเกิดเหตุการณ์ที่สนใจ

ถ้า P (ของ case ใหม่) ≤ 0.5 จะให้เป็น y = 0 หรือไม่เกิดเหตุการณ์ที่สนใจ

0.5 เป็นค่าจุดตัดของความน่าจะเป็นผู้วิเคราะห์อาจกำหนดค่าอื่น ๆ ก็ได้ ตามที่เห็นสมควรของแต่ละเรื่อง เช่น อาจกำหนดเป็น 0.4 หรือ 0.7 ก็ได้ แล้วแต่เห็นเหมาะสม แต่ทั่วไปนิยมใช้ค่า 0.5 เป็นจุดตัด

สรุป

ในการวิเคราะห์การถดถอยโลจิสติค ค่าสัมประสิทธิ์ที่คำนวณได้นำมาเขียนเป็นสมการถดถอยโลจิสติค เพื่อที่จะทำนายความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ ดังนี้

$$P(y = 1) = \frac{1}{1+e^{-z}}$$

ตัวอย่าง การวิจัยเรื่องหนึ่งผู้วิจัยสนใจทำนายการเป็นมะเร็งต่อมน้ำเหลือง สมการถดถอยโลจิสติคที่สร้างได้ประกอบด้วย 2 ตัวแปรทำนาย คือ อายุ และสาร acid โดยที่สมการนี้มีค่าคงที่ ดังนั้นสมการทำนายจึงเป็นดังนี้

$$z = -.337 - .042_{\text{อายุ}} + .076_{\text{สารacid}}$$

จากสมการถดถอยโลจิสติคที่สร้างได้นี้ ถ้านำไปลองทำนาย ชายคนหนึ่งอายุ 62 ปี มีค่าสาร acid เป็น 48 จะมีโอกาสเกิดโรคมะเร็งต่อมน้ำเหลืองหรือไม่

สมการ = - .337 - .042(62) + .076(48)

= -.337 – 2.604 + 3.678

= 0.707

ความน่าเป็นของการเกิดเหตุการณ์ที่สนใจ ในที่นี้คือ การเป็นโรคมะเร็งต่อมน้ำเหลือง P(y = 1)

$$P(y = 1) = \frac{1}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-(0.707)}}$$

$$= \frac{1}{1+2.71828^{-0.707}}$$

$$= 0.493$$

พิจารณาจุดตัด .5 นั่นคือ ถ้ามากกว่า .5 แสดงว่ามีโอกาสเป็นโรคมะเร็งในต่อน้ำเหลือง ในที่นี้มีค่า P = .493 หรือ .5 หรือชายคนนี้มีโอกาสเป็นและไม่ปีนโรคมะเร็งในต่อมน้ำเหลืองเท่ากัน

ข้อเสนอแนะ

การวิเคราะห์การถดถอยโลจิสติคเป็นการวิเคราะห์ที่สนใจ ทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการที่สร้างขึ้นจากข้อมูลเก็บรวบรวมมาที่มีลักษณะแบบตัดขวาง (cross-sectional design) เช่น ผู้วิจัยสนใจโรคมะเร็งปอดเกิดจากการสูบบุหรี่หรือไม่ ก็ไปรวบรวมข้อมูลจากคนเป็นโรคมะเร็งปอดมาจำนวนหนึ่ง จากนั้นจึงพิจารณาเอาผลจากคนเป็นโรคมะเร็งปอดว่าคนที่เป็นโรคมะเร็งปอดสูบหรือไม่สูบบุหรี่ หากพบว่าคนเป็นโรคมะเร็งปอดสูบบุหรี่มีมากกว่าผู้ไม่สูบบุหรี่จะสรุปว่า การสูบบุหรี่ให้ผลการเป็นโรคมะเร็งปอด แต่ก่อนการวิเคราะห์การถดถอยโลจิสติค ผู้วิจัยจะต้องตรวจสอบข้อมูลก่อนว่าเป็นไปตามข้อตกลงหรือไม่

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา. 2549. **การวิเคราะห์สถิติขั้นสูงด้วย SPSS for windows.** พิมพ์ครั้งที่ 5. ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

กัลยา วานิชย์บัญชา. 2551. **การวิเคราะห์ข้อมูลหลายตัวแปร.** พิมพ์ครั้งที่ 3. ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ศิริชัย กาญจนวาสี. 2550. **การวิเคราะห์พหุระดับ = Multi-level analysis.** พิมพ์ครั้งที่ 4. ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ศิริชัย พงษ์วิชัย. 2549. **การวิเคราะห์ข้อมูลทางสถิติด้วยคอมพิวเตอร์.** พิมพ์ครั้งที่ 16. สำนักพิมพ์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

เพชรน้อย สิงห์ช่างชัย. 2549. **หลักการและการใช้สถิติการวิเคราะห์ตัวแปรหลายตัว สำหรับการวิจัยทางการพยาบาล.** พิมพ์ครั้งที่ 3. ชานเมืองการพิมพ์, สงขลา.

Burns, N. and Grove, S.K. 1993. **The Practice of Nursing Research : Conduct, Critique and Utilization.** W.B.Saunders Company, Philadelphia.

Stevens, J. 1996. **Applied multivariate statistics for the social science.** Lawrence Erlbaum Associate, Inc., Mahwah, New Jersey.