# การวิเคราะห์ถดถอยโลจิสติก : แนวคิด การวิเคราะห์ และการแปลความหมาย

## Logistic Regression Analysis : Concept, Analysis and Interpretation

ศิริเดช สุชีวะ


### ABSTRACT

The logistic regression model has been used in statistical analysis for many years; but it was well known after Truett, Cornfield, and Kannel (1967) used the model to provide a multivariate analysis of the Framingham heart study data that its full power and applicability were appreciated.

This article introduced the logistic regression model, its use in methods for modeling the relationship between a dichotomous dependent variable and a set of independent variables as well as the interpretation of the analysis output.

The logistic regression model yielded the product of analysis as same as the discriminant analysis model; but it required the less and more relax assumptions. Thus, the logistic regression model is one of the powerful statistical technique for predicting dichotomous dependent variable.

## บทคัดย่อ

โมเดลการถดถอยโลจิสติก (logictic regression model) เป็นเทคนิคการวิเคราะห์ทางสถิติสำหรับพยากรณ์ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจจากชุดตัวแปรอิสระ โมเดลนี้เป็นที่รู้จักกันดีในหมู่นักสถิติ โดยเริ่มใช้ในการวิจัยทางการแพทย์และสาธารณสุขก่อน และแพร่หลายเข้าสู่การวิจัยในสาขาสังคมศาสตร์ และพฤติกรรมศาสตร์ในภายหลังเมื่อมีโปรแกรมคอมพิวเตอร์ที่สามารถวิเคราะห์โมเดลนี้ได้ เช่น GLIM BMDP SAS และ SPSS สาระที่นำเสนอในบทความนี้ประกอบไปด้วยมโนทัศน์ที่สำคัญของการวิเคราะห์ถดถอยโลจิสติก การเขียนคำสั่ง SPSS PC เพื่อวิเคราะห์ข้อมูล และการแปลความหมายผลการวิเคราะห์ข้อมูล ซึ่งคงจะทำให้ผู้อ่านเข้าใจและสามารถนำเทคนิคการวิเคราะห์นี้ไปใช้ในการวิจัยได้อย่างถูกต้องเหมาะสมยิ่งขึ้น

## 1. แนวคิดของการวิเคราะห์ถดถอยโลจิสติก

เป้าหมายของการวิจัยทางพฤติกรรมศาสตร์ และสังคมศาสตร์ที่สำคัญประการหนึ่งนอกเหนือไปจากการบรรยาย (description) การอธิบาย (explanation) และการควบคุม (control) ก็คือการพยากรณ์ (prediction) ปรากฏการณ์ในธรรมชาติ ซึ่งมีประโยชน์ทั้งในทางวิชาการและในโลกแห่งความเป็นจริง เช่น การพยากรณ์การไปหรือไม่ไปใช้สิทธิออกเสียงเลือกตั้งของประชาชน การพยากรณ์ว่าผู้ใดจะเป็นโรคเส้นเลือดหัวใจหรือไม่เป็น หรือการพยากรณ์ว่าธุรกิจจะล้มเหลว หรือประสบความสำเร็จ เป็นต้น เมื่อนึกถึงเทคนิคการวิเคราะห์ทางสถิติที่ใช้ในการพยากรณ์ตัวแปรตาม จากชุดของตัวแปรอิสระ โดยทั่วไปเราจะนึกถึงการวิเคราะห์ถดถอยพหุ (multiple linear regression analysis, MRA) และการวิเคราะห์จำแนก (discriminant analysis) แต่เทคนิคการวิเคราะห์ทั้งสองแบบนี้ก็ไม่เหมาะที่จะใช้พยากรณ์ตัวแปรตาม ที่มีค่าตัวแปรเพียงสองค่า (ตัวแปรทวิ) คือ การเกิดหรือไม่เกิดเหตุการณ์ เนื่องจากจะเป็นการฝ่าฝืนข้อตกลงเบื้องต้นที่สำคัญของการวิเคราะห์ถดถอยพหุ เช่น ข้อตกลงเบื้องต้นเรื่อง การแจกแจงความคลาดเคลื่อนที่ต้องเป็นโค้งปกติ อีกประการหนึ่ง ค่าที่ทำนายได้จากการวิเคราะห์ถดถอยพหุ ไม่สามารถแปลความหมายเป็นค่าความน่าจะเป็นของการเกิด

เหตุการณ์ได้ โดยค่าที่ทำนายได้นี้อาจจะอยู่นอกช่วง 0 ถึง 1 ได้ ส่วนการวิเคราะห์จำแนกสามารถใช้พยากรณ์การเป็นสมาชิกของกลุ่มได้โดยตรง แต่จะต้องมีข้อตกลงเบื้องต้นในเรื่องการแจกแจงปกติพหุ (multivariate normality) ของตัวแปรอิสระ และเมทริกส์ความแปรปรวน-ความแปรปรวนร่วม (variance-covariance matrix) ที่ต้องเท่ากันในกลุ่มตัวอย่างทั้งสองกลุ่ม ในขณะที่การวิเคราะห์ถดถอยโลจิสติกไม่จำเป็นต้องมีข้อตกลงเบื้องต้นเหล่านี้

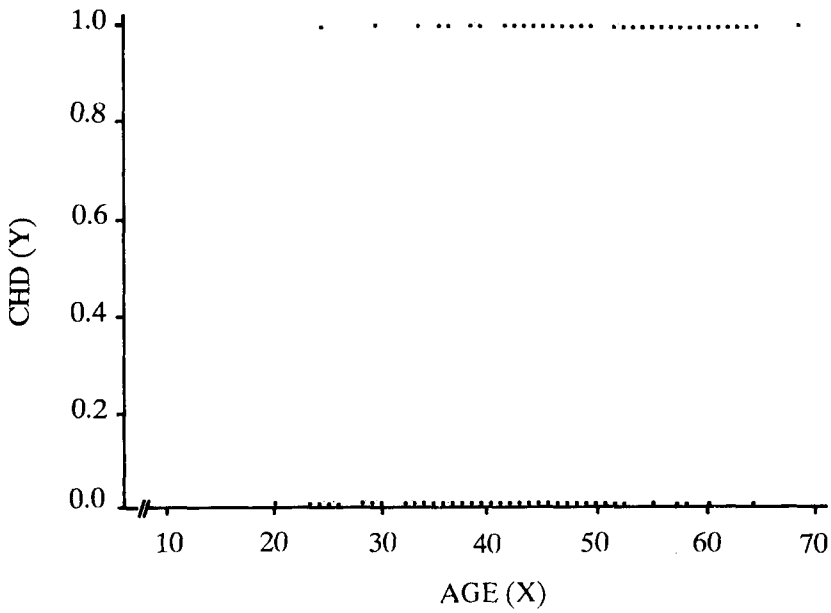
การวิเคราะห์ถดถอยโลจิสติกกับการวิเคราะห์ถดถอยเชิงเส้น แตกต่างกันตรงที่ตัวแปรตามในการวิเคราะห์ถดถอยโลจิสติกจะมี 2 ค่า (binary) หรือเป็นตัวแปร dichotomous ความแตกต่างตรงนี้ทำให้เกิดแนวคิดของการวิเคราะห์ถดถอยโลจิสติก ซึ่งขออธิบายด้วยตัวอย่างการวิจัยเพื่อศึกษาความสัมพันธ์ระหว่างอายุกับการเป็นโรคหลอดเลือดหัวใจต่อไปนี้

ข้อมูลในตารางที่ 1 แสดงค่าตัวแปร 4 ตัว คือ หมายเลขประจำตัว (ID) กลุ่มอายุ (AGRP) อายุ (AGE) และการเป็นโรคหลอดเลือดหัวใจ (CHD) ของกลุ่มตัวอย่าง 100 คน โดย CHD เป็น 0 หมายถึง การไม่เป็นโรคหลอดเลือดหัวใจ และ CHD เป็น 1 หมายถึง การเป็นโรคหลอดเลือดหัวใจ

**ตารางที่ 1** อายุและการเป็นโรคหลอดเลือดหัวใจ (CHD) ของกลุ่มตัวอย่าง 100 คน

| ID | AGRP | AGE | CHD | ID | AGRP | AGE | CHD | ID | AGRP | AGE | CHD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 20 | 0 | 35 | 3 | 38 | 0 | 68 | 6 | 51 | 0 |
| 2 | 1 | 23 | 0 | 36 | 3 | 39 | 0 | 69 | 6 | 52 | 0 |
| 3 | 1 | 24 | 0 | 37 | 3 | 39 | 1 | 70 | 6 | 52 | 1 |
| 4 | 1 | 25 | 0 | 38 | 4 | 40 | 0 | 71 | 6 | 53 | 1 |
| 5 | 1 | 25 | 1 | 39 | 4 | 40 | 1 | 72 | 6 | 53 | 1 |
| 6 | 1 | 26 | 0 | 40 | 4 | 41 | 0 | 73 | 6 | 54 | 1 |
| 7 | 1 | 26 | 0 | 41 | 4 | 41 | 0 | 74 | 7 | 55 | 0 |
| 8 | 1 | 28 | 0 | 42 | 4 | 42 | 0 | 75 | 7 | 55 | 1 |
| 9 | 1 | 28 | 0 | 43 | 4 | 42 | 0 | 76 | 7 | 55 | 1 |
| 10 | 1 | 29 | 0 | 44 | 4 | 42 | 0 | 77 | 7 | 56 | 1 |
| 11 | 2 | 30 | 0 | 45 | 4 | 42 | 1 | 78 | 7 | 56 | 1 |
| 12 | 2 | 30 | 0 | 46 | 4 | 43 | 0 | 79 | 7 | 56 | 1 |
| 13 | 2 | 30 | 0 | 47 | 4 | 43 | 0 | 80 | 7 | 57 | 0 |
| 14 | 2 | 30 | 0 | 48 | 4 | 43 | 1 | 81 | 7 | 57 | 0 |
| 15 | 2 | 30 | 0 | 49 | 4 | 44 | 0 | 82 | 7 | 57 | 1 |
| 16 | 2 | 30 | 1 | 50 | 4 | 44 | 0 | 83 | 7 | 57 | 1 |
| 17 | 2 | 32 | 0 | 51 | 4 | 44 | 1 | 84 | 7 | 57 | 1 |
| 18 | 2 | 32 | 0 | 52 | 4 | 44 | 1 | 85 | 7 | 57 | 1 |
| 19 | 2 | 33 | 0 | 53 | 5 | 45 | 0 | 86 | 7 | 58 | 0 |
| 20 | 2 | 33 | 0 | 54 | 5 | 45 | 1 | 87 | 7 | 58 | 1 |
| 21 | 2 | 34 | 0 | 55 | 5 | 46 | 0 | 88 | 7 | 58 | 1 |
| 22 | 2 | 34 | 0 | 56 | 5 | 46 | 1 | 89 | 7 | 59 | 1 |
| 23 | 2 | 34 | 1 | 57 | 5 | 47 | 0 | 90 | 7 | 59 | 1 |
| 24 | 2 | 34 | 0 | 58 | 5 | 47 | 0 | 91 | 8 | 60 | 0 |
| 25 | 2 | 34 | 0 | 59 | 5 | 47 | 1 | 92 | 8 | 60 | 1 |
| 26 | 3 | 35 | 0 | 60 | 5 | 48 | 0 | 93 | 8 | 61 | 1 |
| 27 | 3 | 35 | 0 | 61 | 5 | 48 | 1 | 94 | 8 | 62 | 1 |
| 28 | 3 | 36 | 0 | 62 | 5 | 48 | 1 | 95 | 8 | 62 | 1 |
| 29 | 3 | 36 | 1 | 63 | 5 | 49 | 0 | 96 | 8 | 63 | 1 |
| 30 | 3 | 36 | 0 | 64 | 5 | 49 | 0 | 97 | 8 | 64 | 0 |
| 31 | 3 | 37 | 0 | 65 | 5 | 49 | 1 | 98 | 8 | 64 | 1 |
| 32 | 3 | 37 | 1 | 66 | 6 | 50 | 0 | 99 | 8 | 65 | 1 |
| 33 | 3 | 37 | 0 | 67 | 6 | 50 | 1 | 100 | 8 | 69 | 1 |
| 34 | 3 | 38 | 0 | | | | | | | | |

เมื่อเป้าหมายสำคัญของการวิจัยอยู่ที่การหาความสัมพันธ์ระหว่างอายุกับการเป็น CHD เราอาจเริ่มต้นโดยการสร้างแผนภาพกระจัดกระจาย (scatterplot) ของตัวแปรตามกับตัวแปรอิสระ ซึ่งจะบอกถึงลักษณะและขนาดของความสัมพันธ์ระหว่างตัวแปร แผนภาพกระจัดกระจายของข้อมูลดังกล่าว แสดงได้ดังรูปที่ 1

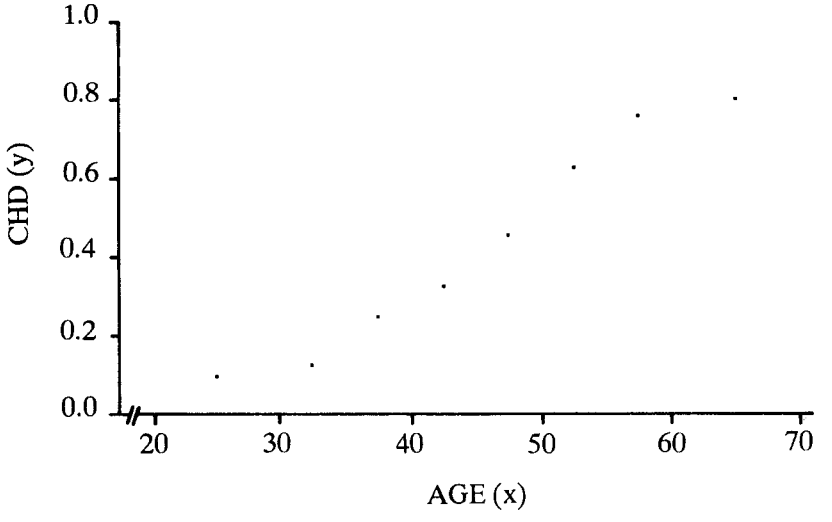



**รูปที่ 1** *แผนภาพกระจัดกระจายของการเป็น CHD และอายุ*

จากแผนภาพกระจัดกระจาย จุดทุกจุดจะตกอยู่บนเส้นขนาน 2 เส้น ที่เป็นตัวแทนของการไม่เป็น CHD (y=0) และการเป็น CHD (y=1) เราไม่สามารถมองเห็นลักษณะความสัมพันธ์ระหว่าง CHD และอายุ ได้อย่างชัดเจน เนื่องมาจากการแปรผันของ CHD ในทุกอายุมีค่อนข้างมาก วิธีการที่จะลดความแปรผันโดยยังคงโครงสร้างของความสัมพันธ์เอาไว้ได้วิธีหนึ่งก็คือการสร้างอันตรภาคของตัวแปรอิสระและคำนวณค่าเฉลี่ยของตัวแปรตามในแต่ละอันตรภาค ดังแสดงในตารางที่ 2 ซึ่งใช้ตัวแปรกลุ่มอายุ (AGRP) มาจัดอันตรภาคของอายุ และใช้สัดส่วนของการเป็น CHD แทนค่า mean ในแต่ละกลุ่มอายุ จากตารางนี้ทำให้มองเห็นภาพความสัมพันธ์ระหว่างอายุกับการเป็น CHD ได้ชัดขึ้นว่า เมื่ออายุเพิ่มขึ้น สัดส่วนของคนเป็น CHD ก็จะเพิ่มขึ้นด้วย

**ตารางที่ 2** *สัดส่วนของการเป็น CHD จำแนกตามกลุ่มอายุ*

| Age Group | n | CHD | | Mean (Proportion) |
|---|---|---|---|---|
| | | Absent | Present | |
| 20-29 | 10 | 9 | 1 | 0.10 |
| 30-34 | 15 | 13 | 2 | 0.13 |
| 35-39 | 12 | 9 | 3 | 0.25 |
| 40-44 | 15 | 10 | 5 | 0.33 |
| 45-49 | 13 | 7 | 6 | 0.46 |
| 50-54 | 8 | 3 | 5 | 0.63 |
| 55-59 | 17 | 4 | 13 | 0.76 |
| 60-69 | 10 | 2 | 8 | 0.80 |
| Total | 100 | 57 | 43 | 0.43 |




**รูปที่ 2** *plot ของสัดส่วนการเป็น CHD ในแต่ละกลุ่มอายุ*

รูปที่ 2 แสดงการ plot สัดส่วนการเป็น CHD กับค่ากลางของแต่ละกลุ่มอายุ ซึ่งดูคล้ายกับ plot ที่ได้จากการวิเคราะห์ถดถอยเชิงเส้น

ข้อแตกต่างสำคัญระหว่างการถดถอยโลจิสติกกับการถดถอยเชิงเส้นมี 2 ประการ ประการแรกเป็นเรื่องลักษณะของความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระในการวิเคราะห์ถดถอยใด ๆ ค่าเฉลี่ยของตัวแปรตาม จะถูกกำหนดโดยค่าของตัวแปรอิสระ ที่เรียกว่าเป็น "ค่าเฉลี่ยแบบมีเงื่อนไข" (conditional mean) ใช้สัญลักษณ์ $E(Y/x)$ โดย Y เป็นตัวแปรตาม และ x เป็นค่าของตัวแปรอิสระ ในการวิเคราะห์ถดถอยเชิงเส้น เราถือว่าค่าเฉลี่ยนี้แสดงได้ในรูปสมการเชิงเส้นของ x

$$E(Y/x) = \beta_0 + \beta_1 x$$

ซึ่งแสดงให้เห็นว่า มีค่า $E(Y/x)$ ที่เป็นไปได้บนค่าของ x ที่มีพิสัย $-\infty$ ถึง $+\infty$

สำหรับข้อมูลที่เป็น dichotomous ค่าเฉลี่ยแบบมีเงื่อนไข จะต้องมีค่าตั้งแต่ 0 ถึง1 ($0 \le E(Y/x) \le 1$) ดังรูปที่ 2 ซึ่งโค้งจะมีลักษณะเป็น S-shaped คล้ายกับรูปของการแจกแจงแบบสะสมของตัวแปรสุ่ม การแจกแจงแบบสะสมที่จะใช้ในโมเดล $E(Y/x)$ ในกรณีที่ Y เป็น dichotomous คือการแจกแจงโลจิสติก (logistic distribution) แม้ว่าจะมีการเสนอฟังก์ชันการแจกแจงแบบอื่นในการวิเคราะห์ตัวแปรตามที่เป็น dichotomous หลายการแจกแจงก็ตาม แต่ คอกซ์ (Cox, 1970, cited in Hosmer and Lemeshow, 1989) ได้อภิปรายไว้ว่าเหตุผลสำคัญ 2 ประการที่เลือกใช้การแจกแจงโลจิสติก คือ ประการแรก ในเชิงคณิตศาสตร์ โลจิสติกเป็นฟังก์ชันที่ค่อนข้างจะยืดหยุ่นมากและใช้ได้ง่าย ประการที่สอง ฟังก์ชันโลจิสติก สามารถแปลความได้อย่างมีความหมายในเชิงชีววิทยาด้วย

เพื่อให้ได้ฟังก์ชันที่ง่ายขึ้น เรากำหนดให้ $\pi(x) = E(Y/x)$ แทนค่าเฉลี่ยแบบมีเงื่อนไขของ Y ที่กำหนดโดย x เมื่อใช้การแจกแจงโลจิสติก โมเดลถดถอยโลจิสติก จึงเขียนได้ดังสมการ

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

การแปลงค่า $\pi(x)$ นี้เรียกว่า การแปลงแบบโลจิท (logit transformation) ซึ่งเขียนในเทอมของ $\pi(x)$ ได้ดังสมการ

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$= \beta_0 + \beta_1 x$$

การแปลงแบบโลจิทนี้ ทำให้ได้ logit หรือ g(x) ที่มีคุณสมบัติเชิงเส้นซึ่งอาจจะมีค่าต่อเนื่องตั้งแต่ $-\infty$ ถึง $+\infty$ ก็ได้ ขึ้นอยู่กับพิสัยของ x

ความแตกต่างระหว่างการวิเคราะห์ถดถอยเชิงเส้นและโลจิสติกที่สำคัญประการที่สองจะเกี่ยวข้องกับการแจกแจงแบบเงื่อนไขของตัวแปรตาม ในการถดถอยเชิงเส้น ค่าของตัวแปรตาม แสดงได้ดังสมการ $Y = E(Y/x)+\varepsilon$ โดยที่ $\varepsilon$ คือความคลาดเคลื่อน ซึ่งมีข้อตกลงเบื้องต้นว่า $\varepsilon$ ต้องมีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็นศูนย์ และมีความแปรปรวนคงที่ในแต่ละระดับของตัวแปรอิสระ แต่ในกรณีที่ตัวแปรตามเป็น dichotomous จะแสดงค่าของตัวแปรตามได้ดังสมการ $Y = \pi(x)+\varepsilon$ โดย $\varepsilon$ มีค่าที่เป็นไปได้ 2 ค่า คือ ถ้า Y เท่ากับ 1 แล้ว $\varepsilon$ จะเท่ากับ $1-\pi(x)$ ด้วยความน่าจะเป็น $\pi(x)$ และถ้า Y เท่ากับ 0 แล้ว $\varepsilon$ จะเท่ากับ $-\pi(x)$ ด้วยความน่าจะเป็น $1-\pi(x)$ ดังนั้น $\varepsilon$ จึงมีการแจกแจงแบบ binomial มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ $\pi(x)[1-\pi(x)]$

กล่าวโดยสรุป แนวคิดในการวิเคราะห์ถดถอยเมื่อตัวแปรตามเป็น dichotomous มี 3 ประการ ได้แก่

(1) จะต้องแปลงค่าเฉลี่ยแบบมีเงื่อนไขของสมการถดถอยให้อยู่ระหว่าง 0 ถึง 1 ซึ่งก็เหมาะสมที่จะใช้โมเดลการถดถอยแบบโลจิสติก

(2) การแจกแจงความคลาดเคลื่อนต้องเป็นแบบ binomial ซึ่งจะเป็นการแจกแจงทางสถิติพื้นฐานในการวิเคราะห์ต่อไป

(3) หลักการอื่น ๆ ของการวิเคราะห์ถดถอยเชิงเส้นสามารถใช้กับการถดถอยโลจิสติกได้ด้วย

## 2. โมเดลการวิเคราะห์ถดถอยโลจิสติก

ในกรณีที่มีตัวแปรอิสระเพียงตัวเดียว โมเดลการวิเคราะห์ถดถอยโลจิสติก สามารถเขียนได้ดังสมการ

$$\text{Prob (event)} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad \text{..... 2.1}$$

หรือ

$$\text{Prob (event)} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad \text{..... 2.2}$$

เมื่อ $\beta_0$ และ $\beta_1$ เป็น สปส. ที่ประมาณได้จากข้อมูล

x เป็นตัวแปรอิสระ

e เป็นล็อกธรรมชาติ (natural logarithms) มีค่าประมาณ 2.718

จากสมการที่ 2.1 และ 2.2 เราสามารถเขียนสมการใหม่ได้เป็น

$$\text{Prob (event)} = \frac{e^Z}{1 + e^Z} \quad \text{..... 2.3}$$

หรือ $$\text{Prob (event)} = \frac{1}{1 + e^{-Z}} \quad \text{..... 2.4}$$

โดย $Z = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p$

และโอกาสของการไม่เกิดเหตุการณ์จะประมาณได้จากสมการ

$$\text{Prob (no event)} = 1 - \text{Prob (event)}$$

ในการวิเคราะห์ถดถอยเชิงเส้น จะประมาณค่าพารามิเตอร์ในโมเดลโดยใช้วิธีกำลังสองน้อยที่สุด โดยคัดเลือก สปส. การถดถอย ที่ทำให้ค่าผลรวมของกำลังสองของความคลาดเคลื่อนในการทำนาย ($\Sigma(Y-\acute{Y})^2$) มีค่าน้อยที่สุด ส่วนการวิเคราะห์ถดถอยโลจิสติก จะประมาณค่าพารามิเตอร์โดยวิธี maximum-likelihood อันเป็นการคำนวณทวนซ้ำ (iterative algorithm) เพื่อให้ได้ค่าประมาณของพารามิเตอร์ที่ใกล้เคียงกับข้อมูลเชิงประจักษ์มากที่สุด

เพื่อให้เห็นภาพของการวิเคราะห์ถดถอยโลจิสติก ผู้เขียนขอเสนอผลการวิจัยของบราวน์ (Brown, 1980 cited in Norusis, 1990) เกี่ยวกับการพยากรณ์การเป็นมะเร็งโดยมีที่มาของปัญหาวิจัยว่า การรักษาและการพยากรณ์การเป็นมะเร็งนั้นขึ้นอยู่กับการลุกลามของเซลมะเร็ง ซึ่งมักเกิดในต่อมน้ำเหลืองก่อน การตรวจสอบมะเร็งในระยะเริ่มแรกจึงต้องอาศัยการผ่าตัดต่อมน้ำเหลือง เพื่อสำรวจว่ามีเซลมะเร็งอยู่หรือไม่ แล้วจึงให้การรักษาที่จำเป็นต่อไป ถ้าเราสามารถพยากรณ์การมีเซลมะเร็งในต่อมน้ำเหลืองได้โดยอาศัยข้อมูลการวินิจฉัยที่มีอยู่ ก็จะไม่ต้องทำการผ่าตัดต่อมน้ำเหลือง และลดค่าใช้จ่ายต่าง ๆ ออกไปได้มาก บราวน์จึงเก็บข้อมูลจากผู้เข้ารับการตรวจมะเร็งระยะเริ่มแรกจำนวน 53 คน โดยเก็บข้อมูล อายุ, serum acid phosphatase (สารที่จะมีค่าสูงขึ้นเมื่อเซลที่เจริญผิดปกติได้แพร่กระจายไปในพื้นที่นั้น), ระยะของโรค (stage), ระดับการผิดปกติของเซล (grade) และผลการ X-ray

รวมทั้งผลการผ่าตัดพิสูจน์เซลมะเร็งในต่อมน้ำเหลือง จากนั้นจึงสร้างสมการถดถอยโลจิสติกเพื่อพยากรณ์การเป็นมะเร็งต่อมน้ำเหลือง จากข้อมูลกลุ่มตัวอย่างทั้ง 53 คน

ในการสร้างสมการถดถอยโลจิสติกเพื่อพยากรณ์ จะใช้คำสั่งดังนี้

```
LOGISTIC REGRESSION VARIABLES = dependent WITH independents.
```

ซึ่งจะให้ได้ผลการวิเคราะห์ข้อมูลดังรูปที่ 3

**Parameter estimates for the logistic regression model**

คำสั่ง:

```
LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE.
```

ผลลัพธ์:

**Variables in the Equation**

| Variable | B | S.E. | Wald | df | Sig | R | Exp(B) |
|---|---|---|---|---|---|---|---|
| AGE | .0693 | .0579 | 1.4320 | 1 | .2314 | .0000 | .9331 |
| ACID | .0243 | .0132 | 3.4229 | 1 | .0643 | .1423 | 1.0246 |
| XRAY | 2.0453 | .8072 | 6.4207 | 1 | .0113 | .2509 | 7.7317 |
| GRADE | .7614 | .7708 | .9758 | 1 | .3232 | .0000 | 2.1413 |
| STAGE | 1.5641 | .7740 | 4.0835 | 1 | .0433 | .1722 | 4.7783 |
| Constant | .0618 | 3.4599 | .0003 | 1 | .9357 | | |

**รูปที่ 3** *ค่าพารามิเตอร์ที่ประมาณได้จากโมเดลถดถอยโลจิสติก*

รูปที่ 3 แสดงค่า สปส. ที่ประมาณได้ (คอลัมน์ B) และค่าสถิติที่เกี่ยวข้องกับการวิเคราะห์ถดถอยโลจิสติกเพื่อทำนายการเป็นมะเร็งต่อมน้ำเหลืองจากตัวแปร AGE, ACID, XRAY, STAGE และ GRADE ตัวแปรอิสระ 3 ตัวสุดท้ายนี้ เป็นตัวแปรบ่งชี้ (indicator variable) มีรหัสเป็น 0 กับ 1 โดย XRAY มีค่าเป็น 1 หมายถึง ให้ผลการ X-ray เป็นบวก STAGE เป็น 1 หมายถึง อยู่ในระยะรุนแรง และ GRADE เป็น 1 หมายถึง การผิดปกติของเซลอยู่ในระดับร้ายแรง

จากค่า สปส. ที่ได้ เราสามารถเขียนสมการถดถอยโลจิสติก ทำนายโอกาสของการเป็นมะเร็งต่อมน้ำเหลืองได้ดังนี้

$$\text{Prob (การเป็นมะเร็งต่อมน้ำเหลือง)} = \frac{1}{1 + e^{-Z}}$$

โดย $$Z = 0.0618 - 0.0693\,(AGE) + 0.0243\,(ACID) + 2.0453\,(XRAY) + 1.5641\,(STAGE) + 0.7614\,(GRADE)$$

ถ้าลองพยากรณ์การเป็นมะเร็งของชายอายุ 66 ปี ที่มีค่า serum acid phosphatase เป็น 48 และค่าตัวแปรที่เหลือเป็น 0 จะได้ว่า

$$Z = 0.0618 - 0.0693\,(66) + 0.0243\,(48) = -3.346$$

$$\text{Prob (การเป็นมะเร็งต่อมน้ำเหลือง)} = \frac{1}{1 + e^{-(-3.346)}} = 0.0340$$

จากผลการวิเคราะห์นี้เราพยากรณ์ว่า เขาไม่น่าจะเป็นมะเร็งต่อมน้ำเหลือง โดยการพยากรณ์ยึดเกณฑ์ความน่าจะเป็น 0.5

## 3. การทดสอบสมมติฐานเกี่ยวกับค่า สปส. ของโมเดลการวิเคราะห์ถดถอยโลจิสติก

การทดสอบสมมติฐานว่า ค่า สปส. ไม่เท่ากับ 0 จะใช้ Wald statistic ซึ่งมีการแจกแจงแบบไค-สแควร์ Wald statistic เป็นกำลังสองของอัตราส่วนระหว่างค่า สปส. กับค่าความคลาดเคลื่อนมาตรฐานของ สปส. นั้น ถ้าเป็นตัวแปรจัดกลุ่ม (categorical variable) Wald statistic มี degree of freedom เท่ากับจำนวนกลุ่มลบด้วย 1 ตัวอย่างเช่น สปส. ของตัวแปร AGE เท่ากับ -0.0693 และความคลาดเคลื่อนมาตรฐานเป็น 0.0579 (ในคอลัมน์ S.E.) Wald statistic จะเท่ากับ $(-0.0693/0.0579)^2$ หรือประมาณ 1.432 นัยสำคัญของ Wald statistic แสดงในคอลัมน์ Sig. จากรูปที่ 3 สปส. ของ XRAY และ STAGE เท่านั้นที่ไม่เท่ากับ 0 อย่างมีนัยสำคัญที่ระดับ 0.05

## 4. สหสัมพันธ์บางส่วน (partial correlation) ในโมเดลการวิเคราะห์ถดถอยโลจิสติก

ในการพิจารณาการมีส่วนร่วมของตัวแปรอิสระแต่ละตัวที่มีต่อการพยากรณ์ ตัวแปรตามนั้นจะดูจากค่าสหสัมพันธ์บางส่วน (partial correlation) ระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัว ดังในคอลัมน์ R statistic ในรูปที่ 3 R มีค่าตั้งแต่ -1 ถึง +1 ค่า R ที่เป็นบวกหมายถึง ถ้าค่าของตัวแปรเพิ่มค่า likelyhood ของการเกิดเหตุการณ์นั้นก็จะเพิ่มขึ้นด้วย ค่า R ที่เป็นลบจะแปลความในทางตรงข้าม

## 5. การแปลความหมายของ สปส. การถดถอย

การวิเคราะห์ถดถอยเชิงเส้นพหุ จะแปลความหมาย สปส. การถดถอยได้โดยตรงว่าเป็นขนาดการเปลี่ยนแปลงของตัวแปรตาม เมื่อตัวแปรอิสระเปลี่ยนไปหนึ่งหน่วย แต่การแปลความหมาย สปส. โลจิสติกจะแตกต่างไปจากนี้ ก่อนอื่นขออธิบายเกี่ยวกับโมเดลโลจิสติกก่อนว่า โมเดลโลจิสติกสามารถเขียนอยู่ในรูปของ odd ของการเกิดเหตุการณ์ได้ (odd ของการเกิดเหตุการณ์ หมายถึง อัตราส่วนระหว่างโอกาสที่จะเกิดกับโอกาสที่จะไม่เกิดเหตุการณ์ เช่น odd ของการออกหัวในการโยนเหรียญ 1 ครั้ง เท่ากับ 0.5/0.5 = 1 เป็นต้น) การเขียนโมเดลโลจิสติกในรูป log ของ odd (ซึ่งเรียกว่า logit) เป็นดังนี้

$$\log\left(\frac{\text{Prob (event)}}{\text{Prob (no event)}}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

จากสมการจะเห็นว่า สปส. โลจิสติก สามารถแปลความได้ว่าเป็นการเปลี่ยนแปลงของ log odd ตามการเปลี่ยนแปลงหนึ่งหน่วยของตัวแปรอิสระ ในรูปที่ 3 สปส. ของ GRADE เป็น 0.76 หมายความว่า ถ้า GRADE เปลี่ยนจาก 0 เป็น 1 และควบคุมตัวแปรอิสระที่เหลือ log odd ของการเป็นมะเร็งต่อมน้ำเหลืองจะเพิ่มขึ้น 0.76 แต่การแปลความหมายในรูปของ odd จะง่ายกว่า log odd ดังนั้น จึงเขียนสมการโลจิสติกใหม่ในเทอมของ odd ได้เป็น

$$\frac{\text{Prob (event)}}{\text{Prob (no event)}} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p} = e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta p X p}$$

e ยกกำลัง $\beta_i$ เป็นค่า odd ที่เปลี่ยนแปลง เมื่อตัวแปรอิสระตัวที่ i มีค่าเพิ่มขึ้น 1 หน่วย ถ้า $\beta_i$ เป็นบวก เทอมนี้จะมีค่ามากกว่า 1 ซึ่งก็หมายความว่า ค่า odd จะเพิ่มขึ้น แต่ถ้า $\beta_i$

เป็นลบ เทอมนี้จะน้อยกว่า 1 หมายความว่า odd จะลดลง ถ้า $\beta_i$ =0 เทอมนี้จะมีค่าเท่ากับ 1 ซึ่งหมายความว่าค่า odd จะไม่เปลี่ยนแปลง เช่น เมื่อ GRADE เปลี่ยนจาก 0 เป็น 1 ค่า odd จะเพิ่มขึ้น 2.14 ดังแสดงในคอลัมน์ Exp ($\beta$) ของรูปที่ 3

ถ้าจะลองคำนวณ odd ของการเป็นมะเร็งต่อมน้ำเหลืองของชายอายุ 60 ปี ที่มี serum acid phosphatase ระดับ 62 ผล X-ray เป็น 1 และผลการวินิจฉัย STAGE และ GRADE เป็น 0 ขั้นแรกต้องคำนวณโอกาสโดยประมาณของการเป็นมะเร็งต่อมน้ำเหลือง

$$\text{Estimated prob (การเป็นมะเร็งต่อมน้ำเหลือง)} = \frac{1}{1 + e^{-Z}}$$

โดย $Z = 0.0618 - 0.0693\,(60) + 0.0243\,(62) + 2.0453\,(1) + 0.7614\,(0) + 1.5641\,(0)$

$= -0.54$

แทนค่า Z ในสมการได้ $\dfrac{1}{1 + e^{-(-0.54)}} = 0.37$

เมื่อโอกาสในการเป็นมะเร็งต่อมน้ำเหลืองเท่ากับ 0.37 และโอกาสที่จะไม่เป็นเท่ากับ 0.63 odd ของการเป็นมะเร็งประมาณได้ดังนี้

$$\text{Odds} = \text{prob (event)/prob (no event)} = \frac{0.37}{1 - 0.37} = 0.59$$

และ log odd เท่ากับ –0.53

## 6. การทดสอบภาวะแนบสนิทดี (goodness of fit) ของโมเดล

การทดสอบความสอดคล้องระหว่างโมเดลกับข้อมูลในการวิเคราะห์ถดถอยโลจิสติก ทำได้หลายวิธี ดังนี้

### 6.1 การใช้ตารางจัดจำพวก (classification table)

วิธีหนึ่งในการทดสอบความสอดคล้องของโมเดล คือการเปรียบเทียบผลการพยากรณ์จากโมเดลกับข้อมูลเชิงประจักษ์ ในโปรแกรม SPSS จะให้ตารางจัดจำพวกออกมา จากคำสั่ง LOGISTIC REGRESSION ดังรูปที่ 4

**Classfication table**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE.

ผลลัพธ์:

**Classification Table for NODES**

| Observed | | Negative N | Positive P | Percent Correct |
|---|---|---|---|---|
| Negative | N | 28 | 5 | 84.85% |
| Positive | P | 7 | 13 | 65.00% |
| | | | Overall | 77.36% |

**รูปที่ 4** *ตารางจัดจำพวก*

จากตาราง ผู้ป่วยที่ไม่ได้เป็นมะเร็ง 28 คน ได้รับการพยากรณ์อย่างถูกต้องจากโมเดลว่าไม่เป็นมะเร็ง เช่นเดียวกับผู้ที่เป็นมะเร็ง 13 คน ก็ได้รับการพยากรณ์จากโมเดลอย่างถูกต้องเช่นเดียวกัน ส่วนในอีก 2 เซลที่เหลือเป็นจำนวนของผู้ที่ได้รับการพยากรณ์ผิดจำนวนทั้งสิ้น 12 คน กล่าวได้ว่า โมเดลนี้พยากรณ์ผู้ที่ไม่เป็นมะเร็งได้ถูกต้อง 84.85% พยากรณ์ผู้ที่เป็นมะเร็งได้ถูกต้อง 65% เมื่อพิจารณาในภาพรวม ถือว่าพยากรณ์ได้ถูกต้อง 77.63% จากผู้ป่วย 53 คน

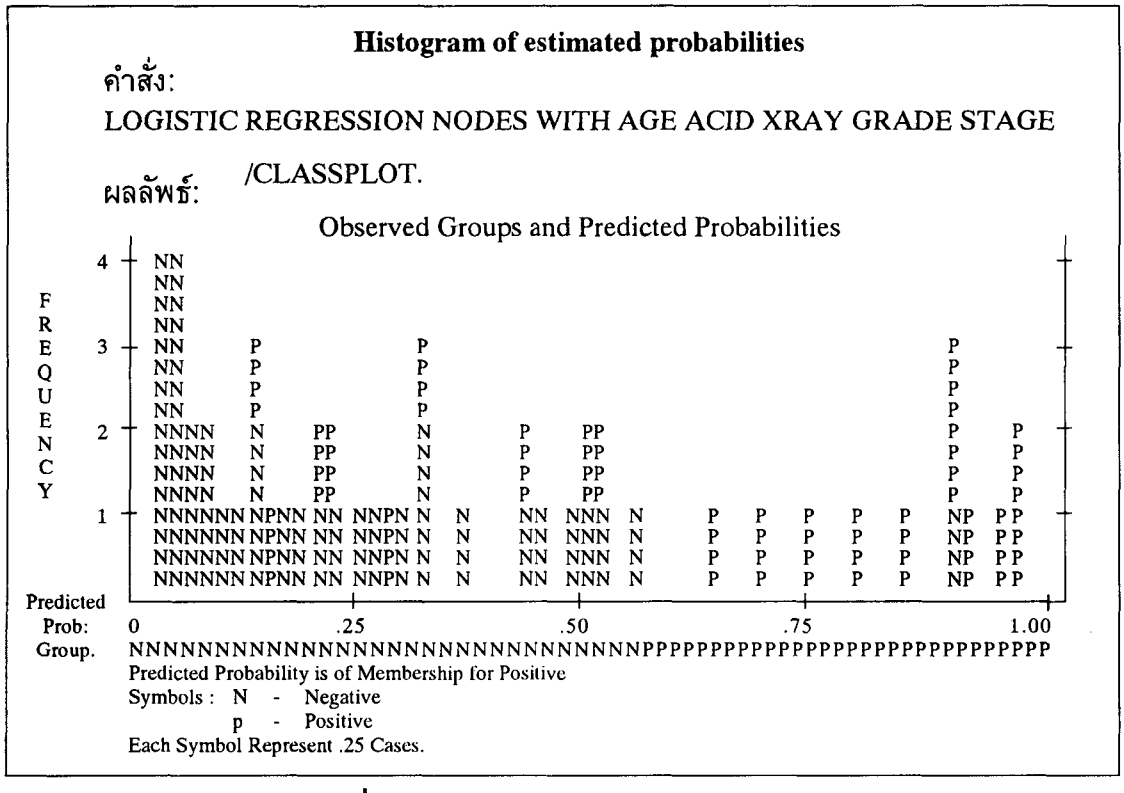
ตารางจัดจำพวกไม่ได้แสดงให้เห็นการแจกแจงความน่าจะเป็นที่ประมาณได้สำหรับผู้ป่วยแต่ละคน เราจึงไม่สามารถบอกได้ว่าผู้ป่วยที่ได้รับการพยากรณ์ผิดว่าไม่เป็นมะเร็งจำนวน 7 คน มีค่าความน่าจะเป็นมะเร็งเข้าใกล้ 0.05 หรือต่ำกว่านี้เพียงใด ซึ่งการทราบข้อมูลตรงนี้จะช่วยในการตัดสินใจได้ดีขึ้น อันเป็นที่มาของวิธีการทดสอบแบบที่ 2

### 6.2 ฮีสโตแกรมของค่าประมาณความน่าจะเป็น (Histogram of Estimated Probabilities)

การทดสอบความเหมาะสมของโมเดลด้วยวิธีนี้ สามารถสั่งได้โดยใช้คำสั่งรอง /CLASSPLOT ดังนี้

LOGISTIC REGRESSION dependent variable WITH independent variables
/CLASSPLOT.

ผลการวิเคราะห์เป็นดังรูปที่ 5




**รูปที่ 5** *ฮีสโตแกรมของค่าประมาณความน่าจะเป็น*

ในรูปที่ 5 สัญลักษณ์ N แทนผู้ไม่เป็นมะเร็ง และ P แทนผู้ที่เป็นมะเร็ง โดย N และ P แต่ละตัวจะแทนคน 0.25 คน ถ้าโมเดลสามารถพยากรณ์คนสองกลุ่มนี้ได้อย่างถูกต้องโดยสมบูรณ์ คนที่เป็นมะเร็งจะอยู่ทางด้านขวาของค่า 0.5 บนแกนนอน ส่วนคนที่ไม่เป็นมะเร็งจะอยู่ทางด้านซ้ายทั้งหมด จากรูปจะเห็นได้ว่า มีคนที่ไม่เป็นมะเร็งเพียงคนเดียวที่มีค่าประมาณโอกาสของการเป็นมะเร็งสูง (ประมาณ 0.88) และมีผู้ที่เป็นมะเร็ง 4 คนที่มีค่าประมาณโอกาสของการเป็นมะเร็งต่ำกว่า 0.25 การพิจารณาฮีสโตแกรมของโอกาสที่ทำนายได้จะเป็นประโยชน์ต่อการตัดสินใจ เช่น ถ้าการพยากรณ์ผิดส่วนใหญ่อยู่ใกล้ ๆ กับเขตความน่าจะเป็น 0.5 เราอาจจะไม่ตัดสินคนที่มีค่าความน่าจะเป็นใกล้ ๆ 0.5 ก็ได้ ซึ่งหมายความว่าเราจะตัดสินการเป็นมะเร็งจากโมเดลเฉพาะรายที่มั่นใจว่าโมเดลโลจิสติกจะทำนายได้ถูกต้องจริง ๆ เท่านั้น ส่วนรายที่ยังไม่มั่นใจก็อาจใช้วิธีผ่าตัดทั้งหมด

### 6.3 ภาวะแนบสนิทดีของโมเดล (goodness of fit of the model)

การทดสอบความเหมาะสมของโมเดลอีกวิธีหนึ่ง คือ การพิจารณาว่า ข้อมูลเชิงประจักษ์จากกลุ่มตัวอย่างใกล้เคียงกับค่าพารามิเตอร์ที่ประมาณได้เพียงใด โอกาสซึ่งผลที่สังเกตได้จะให้ค่าประมาณพารามิเตอร์ เรียกว่า likelihood เนื่องจาก likelihood เป็นตัวเลขที่มีค่าน้อยกว่า 1 จึงมักใช้ -2 คูณกับ log ของ likelihood (-2 LL) ในการทดสอบว่าโมเดลสอดคล้องกับข้อมูลเพียงใด โมเดลที่ดีจะมีค่า likelihood ของผลที่สังเกตได้สูง (ถ้าโมเดลสอดคล้องกับข้อมูลอย่างสมบูรณ์ ค่า likelihood จะเป็น 1 และ -2 คูณกับ log ของ likelihood จะเท่ากับ 0) ผลการทดสอบด้วยวิธีนี้จะออกมาจากคำสั่ง LOGISTIC REGRESSION ดังรูปที่ 6

**2 LL for model containing only the constant**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE.

ผลลัพธ์:

Dependent Variable NODES

Beginning Block Number 0 Initial Log Likelihood Function

–2 Log Likelihood 70.252153

* Constant is included in the model.

**รูปที่ 6** *-2 log likelihood ของโมเดลที่มีเฉพาะค่าคงที่*

ในโมเดลถดถอยโลจิสติกที่มีเฉพาะค่าคงที่ -2 LL จะเท่ากับ 70.25 ตามรูปที่ 6 ส่วนผลการวิเคราะห์ของโมเดลที่มีตัวแปรอิสระครบทุกตัวแสดงในรูปที่ 7

ในการทดสอบสมมติฐานว่าค่า likelihood ที่สังเกตได้ ไม่แตกต่างจาก 1 ซึ่งเป็นค่า likelihood ของโมเดลที่สอดคล้องกับข้อมูลอย่างสมบูรณ์ เราจะใช้ค่า -2 LL ภายใต้สมมติฐานศูนย์ว่า โมเดลสอดคล้องกับข้อมูลอย่างสมบูรณ์ -2 LL จะมีการแจกแจงแบบไค-สแควร์ มี degree of freedom เป็น N-P เมื่อ N เป็นจำนวนกลุ่มตัวอย่าง และ P เป็นจำนวนค่าพารามิเตอร์ที่จะประมาณ ในกรณีนี้ N เท่ากับ 53 และ P เท่ากับ 6 degree of freedom จึงเป็น 47 ผลการทดสอบ ไม่สามารถปฏิเสธสมมติฐานศูนย์ได้

สถิติอีกชนิดหนึ่งที่ใช้ทดสอบความเหมาะสมของโมเดล คือ goodness-of-fit statistic ซึ่งจะเปรียบเทียบความน่าจะเป็นที่สังเกตได้กับความน่าจะเป็นที่ทำนายได้จากโมเดลดังสมการ

$$Z^2 = \sum \frac{\text{Residual}_i^2}{P_i(1-P_i)}$$

residual เป็นค่าความแตกต่างระหว่างค่าที่สังเกตได้ ($Y_i$) และค่าที่ทำนายได้ ($P_i$) สถิติชนิดนี้มีการแจกแจงแบบไค-สแควร์เช่นเดียวกัน ภายใต้สมมติฐานว่าโมเดลสอดคล้องกับข้อมูล และ degree of freedom เป็น N-P ผลการทดสอบด้วย goodness-of-fit statistic ให้ข้อสรุปเหมือนกับ -2 LL ดังในรูปที่ 7

รูปที่ 7 แสดง goodness-of-fit statistics ของโมเดลกับตัวแปรอิสระทั้งหมด แถวแรกของตารางเป็นการเปรียบเทียบโมเดลปัจจุบันกับโมเดลสมบูรณ์แบบ ค่าไค-สแควร์ของ -2LL ในโมเดลปัจจุบันเท่ากับ 48.126 ซึ่งผลการทดสอบพบว่าไม่แตกต่างอย่างมีนัยสำคัญ จากโมเดลสมบูรณ์แบบ และ goodness-of-fit statistic ในแถวสุดท้ายก็ให้ผลเช่นเดียวกัน

**Statistics for model containing the inependent variables**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE.

ผลลัพธ์:

| | Chi-Square | df | Significance |
|---|---|---|---|
| 2 Log Likelihood | 48.126 | 47 | .4270 |
| Model Chi-Square | 22.126 | 5 | .0005 |
| Improvement | 22.126 | 5 | .0005 |
| Goodness of Fit | 46.790 | 47 | .4812 |

**รูปที่ 7** *ค่าสถิติสำหรับโมเดลที่มีตัวแปรอิสระ*

สิ่งที่เพิ่มมาในรูปที่ 7 คือ **Model Chi-Square** และ **Improvement** ในตัวอย่างนี้ model chi-square เป็นความแตกต่างระหว่าง -2 LL ของโมเดลที่มีแต่ค่าคงที่ และ -2 LL ของโมเดลปัจจุบัน ดังนั้น model chi-square จึงทดสอบสมมติฐานสูญที่ว่า สปส. ของทุกเทอมในโมเดลปัจจุบัน ยกเว้นค่าคงที่ มีค่าเป็นศูนย์ การทดสอบตรงนี้เปรียบได้กับการทดสอบ overall F test ในการวิเคราะห์ถดถอยเชิงเส้น ในตัวอย่างนี้ -2 LL ของโมเดลที่มีแต่ค่าคงที่

เป็น 70.25 (จากรูปที่ 6) ขณะที่โมเดลสมบูรณ์แบบจะเป็น 48.126 model chi-square จึงเท่ากับ 22.126 degree of freedom ของ model chi-square เท่ากับความแตกต่างระหว่าง degree of freedom ของสองโมเดลที่เปรียบเทียบกัน กรณีนี้ degree of freedom ของโมเดลที่มีค่าคงที่ เป็น 52 ส่วนโมเดลที่มีค่าคงที่และตัวแปรอิสระ 5 ตัว มี degree of freedom เป็น 47 model chi-square จึงมี degree of freedom เป็น 5 ส่วนแถวที่มีชื่อ **Improvement** เป็นการเปลี่ยนแปลงของ -2 LL ระหว่างขั้นตอนในการสร้างโมเดล ใช้ทดสอบสมมติฐานศูนย์ที่ว่า สปส. ของตัวแปรที่เพิ่มเข้ามาในขั้นตอนสุดท้ายป็น 0 ในกรณีการวิจัยนี้ เราพิจารณาเฉพาะ 2 โมเดล คือ โมเดลที่มีแต่ค่าคงที่กับโมเดลที่มีค่าคงที่และตัวแปรอิสระ 5 ตัว ดังนั้น model chi-square และ improvement chi-square จึงมีค่าเท่ากัน ถ้าเราใช้การคัดเลือกตัวแปรแบบ forward หรือ backward จะทำให้ model chi-square กับ improvement chi-square จะมีค่าแตกต่างกัน การทดสอบ improvement chi-square เปรียบได้กับการทดสอบ F-change ในการวิเคราะห์ถดถอยเชิงเส้นนั่นเอง

## 7. การคัดเลือกตัวแปรพยากรณ์

ในการวิเคราะห์ถดถอยโลจิสติก เราสามารถัดเลือกตัวแปรพยากรณ์ที่ดีได้เช่นเดียวกับการวิเคราะห์ถดถอยเชิงเส้น ในคำสั่ง LOGISTIC REGRESSION เราอาจใช้คำสั่งรอง /ENTER เพื่อให้ตัวแปรทุกตัวเข้าสู่สมการ หรืออาจใช้คำสั่งรอง/FSTEP. สำหรับคัดเลือกแบบ forward stepwise หรือ /BSTEP สำหรับคัดเลือกแบบ backward stepwise ก็ได้

### 7.1 การคัดเลือกตัวแปรพยากรณ์แบบ forward stepwise

การคัดเลือกตัวแปรพยากรณ์แบบ forward stepwise มีวิธีการดำเนินงานเหมือนการคัดเลือกในการวิเคราะห์ถดถอยเชิงเส้นพหุ เริ่มจากโมเดลที่มีแต่ค่าคงที่ ในแต่ละขั้น ตัวแปรที่มีคะแนนที่มีระดับนัยสำคัญน้อยที่สุด จะถูกคัดเข้าสู่โมเดล ตัวแปรทุกตัวที่เข้ามาอยู่ในโมเดลแล้วจะถูกพิจารณาว่าอยู่ในเกณฑ์ที่จะคัดออกหรือไม่ โดยปกติใช้ Wald statistic ในการคัดตัวแปรออก ตัวแปรที่มีระดับนัยสำคัญของ Wald statistic มากที่สุด เกินกว่าจุดตัดที่กำหนดไว้ (โดยปกติเป็น 0.1) จะถูกคัดออกจากโมเดล ถ้าไม่มีตัวแปรใดอยู่ในเกณฑ์ที่จะคัดออกแล้วก็จะคัดเลือกตัวแปรต่อไปเข้าสู่โมเดล แล้วพิจารณาคัดเลือกตัวแปรในโมเดลใหม่นี้ออกอีก ถ้ามีตัวแปรใดถึงเกณฑ์ที่จะถูกคัดออก วนไปเช่นนี้จนกว่าจะไม่มีตัวแปรที่อยู่ในเกณฑ์คัดเลือกเข้าหรือคัดเลือกออกแล้ว แต่เกณฑ์ที่ดีกว่าในการคัดตัวแปรออกจากโมเดล คือ การทดสอบ

อัตราส่วน likelihood หรือ likelihood ration (LR) ซึ่งคำนวณซับซ้อนกว่า Wald statistic วิธีการนี้จะพิจารณาการเปลี่ยนแปลงของ log likelihood เมื่อตัวแปรแต่ละตัวถูกคัดออกจากโมเดล LR ใช้ทดสอบสมมติฐานศูนย์ว่า สปส. ของเทอมที่คัดออกเป็นศูนย์ หาได้จากการหาร likelihood ของโมเดลสมบูรณ์แบบด้วย likelihood ของโมเดลที่มีการคัดตัวแปรออกแล้ว หากสมมติฐานศูนย์เป็นจริงและขนาดของกลุ่มตัวอย่างใหญ่พอ ค่า -2 คูณกับ log ของ LR จะแจกแจงแบบไค-สแควร์ มี degree of freedom เท่ากับ r ซึ่งเป็นความแตกต่างระหว่างจำนวนเทอมในโมเดลเต็มรูปกับโมเดลลดรูป (ทั้ง model chi-square และ improvement chi-square ต่างก็เป็นการทดสอบ LR ด้วย) ในการคัดเลือกตัวแปรด้วยวิธีนี้ จะเปรียบเทียบระดับนัยสำคัญของ LR กับค่าจุดตัด ขั้นตอนการคัดเลือกก็เหมือนกับที่กล่าวมาแล้ว เพียงแต่จะใช้สถิติ LR แทน Wald statistic ในการคัดตัวแปรออก ลองพิจารณาผลของการคัดเลือกตัวแปรแบบ forward stepwise ด้วยคำสั่ง

```
LOGISTIC REGRESSION dependent variable WITH independent variables
/FSTEP.
```

ซึ่งให้ผลการวิเคราะห์ดังรูปที่ 8

**Variables not in the equation**

คำสั่ง:

```
LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
/FSTEP.
```

ผลลัพธ์:

**Variables in the Equation**

| Variable | B | S.E. | Wald | df | Sig | R | Exp (B) |
|---|---|---|---|---|---|---|---|
| Constant | .5008 | .2834 | 3.1227 | 1 | .0772 | | |

**Variables not in the Equation**

Residual Chi Square 19.451 with 5 df Sig = .0016

| Variable | Score | df | Sig | R |
|---|---|---|---|---|
| AGE | 1.0945 | 1 | .2955 | .0000 |
| ACID | 3.1168 | 1 | .0775 | .1261 |
| XRAY | 11.2829 | 1 | .0008 | .3635 |
| GRADE | 4.0745 | 1 | .0435 | .1718 |
| STAGE | 7.4381 | 1 | .0064 | .2782 |

**รูปที่ 8** *ค่าสถิติของตัวแปรที่ไม่อยู่ในสมการ*

จากรูปแสดงให้เห็นค่าสถิติในโมเดลเมื่อค่าคงที่เข้าสู่สมการเพียงตัวเดียว ในขั้นแรกเราจะดูค่าสถิติของค่าคงที่ แล้วดูค่าสถิติของตัวแปรที่ไม่ได้อยู่ในสมการ (ค่า R ของตัวแปรที่ไม่ได้อยู่ในสมการ คำนวณจาก score statistic แทนที่จะเป็น Wald statistic)

ในบรรทัดที่ 5 residual chi-square statistic ทดสอบสมมติฐานสูญว่า สปส. ของทุกตัวแปรที่ไม่ได้อยู่ในโมเดลมีค่าเป็นศูนย์ ถ้าระดับนัยสำคัญของ residual chi-square statistic มีค่าน้อย (ทำให้ปฏิเสธสมมติฐานสูญได้) จึงจะสมควรคัดเลือกตัวแปรต่อ แต่ถ้าไม่สามารถปฏิเสธสมมติฐานศูนย์ว่า สปส. ทุกตัวเท่ากับศูนย์ได้ ก็ควรหยุดการคัดเลือก ถ้ายังสร้างโมเดลต่อไป ก็มีโอกาสว่าโมเดลที่สร้างขึ้นจะใช้ไม่ได้กับกลุ่มตัวอย่างกลุ่มอื่นในประชากรเดียวกันนี้ จากตัวอย่างผลการวิเคราะห์ ระดับนัยสำคัญของ residual chi-square มีค่าน้อย จึงดำเนินการคัดเลือกต่อไปได้ จากรูปตัวแปรแต่ละตัวจะแสดงค่า score statistic (Rao's efficient score statistic) และระดับนัยสำคัญ ถ้าตัวแปรถูกคัดเลือกเข้าสู่โมเดล score statistic จะใช้ทดสอบสมมติฐานสูญที่ว่า สปส. เท่ากับศูนย์ได้มีประสิทธิภาพกว่า Wald statistic โดย LR statistic, Wald statistic และ Rao's efficient score statistic จะมีประสิทธิภาพเท่ากันในกลุ่มตัวอย่างขนาดใหญ่ เมื่อสมมติฐานสูญเป็นจริงเท่านั้น (Rao, 1973 cited in Norusis, 1990)

จากรูปที่ 8 XRAY มี score statistic ขนาดใหญ่ที่สุดและมีระดับนัยสำคัญต่ำกว่า 0.05 จึงถูกคัดเข้าสู่โมเดล ค่าสถิติของตัวแปรที่อยู่นอกโมเดลในตอนนี้แสดงในรูปที่ 9 ซึ่งจะเห็นว่าตัวแปร STAGE มี score statistic ขนาดใหญ่ที่สุด และมีนัยสำคัญ จึงถูกคัดเข้าสู่โมเดลเป็นตัวต่อมา รูปที่ 10 แสดงค่า สปส. โลจิสติก เมื่อ STAGE เข้าสู่โมเดลแล้ว เนื่องจากระดับนัยสำคัญของสปส. ของตัวแปรทั้งสองตัวน้อยกว่า 0.1 ซึ่งเป็นเกณฑ์คัดออก จึงไม่มีตัวแปรใดออกจากโมเดล

**Variables not in the equation**

คำสั่ง:

```
LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
        /FSTEP.
```

ผลลัพธ์:

**Variables not in the Equation**

Residual Chi Square 10.360 with 4 df Sig = .0348

| Variable | Score | df | Sig | R |
|---|---|---|---|---|
| AGE | 1.3524 | 1 | .2449 | .0000 |
| ACID | 2.0732 | 1 | .1499 | .0323 |
| GRADE | 2.3710 | 1 | .1236 | .0727 |
| STAGE | 5.6393 | 1 | .0176 | .2276 |

**รูปที่ 9** *ค่าสถิติของตัวแปรที่ไม่ได้อยู่ในสมการ*

**Logistic coefficients with STAGE and XRAY**

คำสั่ง:

```
LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
        /FSTEP.
```

ผลลัพธ์:

**Variables in the Equation**

| Variable | B | S.E. | Wald | df | Sig | R | Exp (B) |
|---|---|---|---|---|---|---|---|
| XRAY | 2.1194 | .7468 | 8.0537 | 1 | .0045 | .2935 | 8.3265 |
| STAGE | 1.5883 | .7000 | 5.1479 | 1 | .0233 | .2117 | 4.8953 |
| Constant | -2.0446 | .6100 | 11.2360 | 1 | .0008 | | |

**รูปที่ 10** *สปส. โลจิสติกของ STAGE และ XRAY*

goodness-of-fit statistic ของโมเดลที่มีตัวแปร XRAY และ STAGE แสดงไว้ในรูปที่ 11 ทั้งการทดสอบ -2 LL และ goodness-of-fit บ่งชี้ว่าโมเดลสอดคล้องกับข้อมูล model chi-square หรือความแตกต่างระหว่าง -2 LL เมื่อมีเพียงค่าคงที่ในโมเดล กับ -2 LL เมื่อมีทั้งค่าคงที่, XRAY และ STAGE อยู่ในโมเดล (70.25-53.35=16.90) ชี้ให้เห็นว่ามีการเปลี่ยนแปลงอย่างมีนัยสำคัญในโมเดล improvement chi-square เป็นความแตกต่างใน -2 LL เมื่อ STAGE ถูกคัดเข้าสู่โมเดลร่วมกับ XRAY และค่าคงที่ จากระดับนัยสำคัญที่วิเคราะห์ได้ แสดงให้เห็นว่าการนำ STAGE เข้าสู่โมเดล ทำให้โมเดลดีขึ้นอย่างมีนัยสำคัญ -2 LL ของโมเดลที่มีค่าคงที่และ XRAY เป็น 59.001 ดังนั้น improvement chi-square จึงมีค่าเท่ากับ 59.00-53.35=5.65

**Goodness-of-fit statistics with STAGE and XRAY**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
/FSTEP.

ผลลัพธ์:

| | Chi-square | df | Significance |
|---|---|---|---|
| 2 Log Likelihood | 53.353 | 50 | .3466 |
| Model Chi-Square | 16.899 | 2 | .0002 |
| Improvement | 5.647 | 1 | .0175 |
| Goodness of Fit | 54.018 | 50 | .3235 |

**รูปที่ 11** *Goodness-of-fit statistics เมื่อ STAGE และ XRAY เข้าสู่สมการ*

ค่าสถิติของตัวแปรที่อยู่นอกโมเดล หลังจากที่ STAGE ได้เข้าไปแล้ว แสดงไว้ดังรูปที่ 12 ซึ่งจะเห็นได้ว่าระดับนัยสำคัญทั้งสามมีค่าสูงกว่า 0.05 จึงไม่มีการนำตัวแปรเข้าสู่โมเดลอีก

**Variables not in the model after STAGE**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
/FSTEP.

ผลลัพธ์:

Variables not in the Equation

| Residual Chi Square | | 5.422 | with | 3 df | Sig = .1434 |
|---|---|---|---|---|---|
| Variable | Score | df | Sig | R | |
| AGE | 1.2678 | 1 | .2602 | .0000 | |
| ACID | 3.0917 | 1 | .0787 | .1247 | |
| ORADE | .5839 | 1 | .4448 | .0000 | |

**รูปที่ 12** *ค่าสถิติของตัวแปรที่ยังไม่ได้เข้าสู่สมการเมื่อ STAGE ได้เข้าไปแล้ว*

ถ้าเราเลือกใช้ likelihood-ration statistic (LR) แทน Wald statistic ในการดึงตัวแปรออกจากสมการ โดยใช้คำสั่งรอง/FSTEP (LR) ผลการวิเคราะห์จะแตกต่างจากข้างต้นเล็กน้อย สำหรับตัวแปรในสมการ ณ ขั้นตอนหนึ่ง ๆ จะแสดงผลการวิเคราะห์ดังรูปที่ 13 เพิ่มเติมไปจาก สปส. ปกติ และ Wald statistic

**Variables not in the model after STAGE**

คำสั่ง:

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE
/FSTEP (LR).

ผลลัพธ์:

Model if Term Removed

| Term Removed | Log Likelihood | -2 Log LR | df | Significance of Log LR |
|---|---|---|---|---|
| XRAY | -31.276 | 9.199 | 1 | .0024 |
| STAGE | -29.500 | 5.647 | 1 | .0175 |

**รูปที่ 13** *ค่าสถิติของตัวแปรที่ยังไม่ได้เข้าสู่สมการเมื่อ STAGE ได้เข้าไปแล้ว (/FSTEP LR)*

ในรูปที่ 13 ในตัวแปรแต่ละตัวจะแสดงค่า log likelihood เมื่อสมมติว่าตัวแปรตัวนั้นถูกคัดออกจากโมเดล และค่า -2 LR ซึ่งใช้ทดสอบสมมติฐานสูญที่ว่า สปส. ของเทอมเป็นศูนย์ ถ้าระดับนัยสำคัญสูงกว่าจุดตัดของการคงอยู่ในโมเดล ก็จะคัดตัวแปรนั้นออกจากโมเดล และค่าสถิติต่าง ๆ จะถูกคำนวณใหม่ เพื่อดูว่ามีตัวแปรใดที่ควรถูกคัดออกอีกหรือไม่

### 7.2 การคัดออกโดยวิธี backward stepwise

การคัดเลือกตัวแปรแบบ forward เริ่มต้นจากการไม่มีตัวแปรในโมเดล ในขณะที่การคัดเลือกแบบ backward เริ่มต้นโดยนำตัวแปรทุกตัวเข้าสู่โมเดล ในแต่ละขั้นตอนตัวแปรจะถูกพิจารณาคัดเข้าและคัดออก โดยใช้ score statistic ในการตัดสินว่า ตัวแปรควรจะถูกคัดเข้าสู่โมเดลหรือไม่ ส่วน Wald statistic หรือ Likelihood-ration statistic ใช้ในการคัดตัวแปรออกจากสมการเช่นเดียวกับวิธี forward

## 8. บทสรุป

จากที่กล่าวมาทั้งหมด คงจะทำให้ผู้อ่านมองเห็นภาพของการวิเคราะห์ถดถอยโลจิสติกและการแปลความหมายผลการวิเคราะห์ได้ชัดเจนขึ้น การเลือกใช้สถิติให้เหมาะสมกับเป้าหมายของการวิจัยเป็นเรื่องที่สำคัญสำหรับการสร้างสรรค์งานวิจัยที่มีคุณภาพ การวิเคราะห์ถดถอยโลจิสติก เป็นทางเลือกหนึ่งสำหรับการวิจัยที่มีเป้าหมายเพื่อพยากรณ์ตัวแปรตามที่มีลักษณะเป็น dichotomous โดยอาศัยหลักการของการวิเคราะห์ถดถอยเชิงเส้น แต่มีการแปลความหมายสัมประสิทธิ์การถดถอย ที่แตกต่างออกไป โดยต้องแปลความหมายในรูปของอัตราส่วนแต้มต่อ (odd ratio) และโมเดลถดถอยโลจิสติกสามารถใช้วิเคราะห์ข้อมูลที่มีลักษณะตามข้อตกลงเบื้องต้นของการวิเคราะห์จำแนก (discriminant analysis) ได้เป็นอย่างดีด้วย (Hosmer, D.W. and Lemeshow, S., 1989) การวิเคราะห์ถดถอยโลจิสติกจึงเป็นทางเลือกที่น่าสนใจอีกทางหนึ่ง

## เอกสารอ้างอิง

Cox, D.R. (1970). ***The Analysis of Binary Data.*** London : Methuen.

Hosmer, D.W., and Lemeshow, S. (1989). ***Applied Logistic Regression.*** New York : John Wiley & Sons.

Norusis, Marija J. (1990). ***SPSS/PC : Advanced Statistic.*** Chicago : SPSS inc.

Rao, C.R. (1973). ***Linear Statistic Inference and Its Application***. Second Edition. New York : John Wiley & Sons.

Truett, J., Cornfield, J., and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. ***Journal of Chronic Disease***, 20. 511-524.