วิทยานิพนธ์

การศึกษาเปรียบเทียบการวินิจฉัยโรคตับด้วยการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก กรณีศึกษาผู้ป่วย โรงพยาบาลเมืองฉะเชิงเทรา

A COMPARATIVE STUDY OF LIVER DISEASE DIAGNOSIS USING DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION ANALYSIS FOR THE PATIENTS IN CHACHOENGSAO HOSPITAL

นางสาวศยามล ลำลองรัตน์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2550

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (สถิติ)

ปริญญา

สถิติ — **สาขา**

สถิติ — **ภาควิชา**

เรื่อง การศึกษาเปรียบเทียบการวินิจฉัยโรคตับด้วยการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก กรณีศึกษาผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา

A Comparative Study of Liver Disease Diagnosis Using Discriminant Analysis and Logistic Regression Analysis for the Patients in Chachoengsao Hospital

นามผู้วิจัย นางสาวศยามล ลำลองรัตน์

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(ผู้ช่วยศาสตราจารย์บุญอ้อม โฉมที, Ph.D.)

กรรมการ .....

(รองศาสตราจารย์อนันต์ชัย เชื่อนธรรม, M.S.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์สาโรจน์ โอพิทักษ์ชีวิน, M.B.A.)

หัวหน้าภาควิชา .....

(รองศาสตราจารย์สายสุดา สมชิต, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การศึกษาเปรียบเทียบการวินิจฉัยโรคตับด้วยการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก กรณีศึกษาผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา

A Comparative Study of Liver Disease Diagnosis Using Discriminant Analysis and Logistic Regression Analysis for the Patients in Chachoengsao Hospital

โดย

นางสาวศยามล ลำลองรัตน์

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)
พ.ศ. 2550

ศยามล ลำลองรัตน์ 2550: การศึกษาเปรียบเทียบการวินิจฉัยโรคตับด้วยการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก กรณีศึกษาผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ ประธานกรรมการที่ปรึกษา: ผู้ช่วยศาสตราจารย์บุญอ้อม โฉมที, Ph.D. 91 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการวินิจฉัยการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราที่ได้รับการตรวจการทำงานของตับซึ่งแบ่งเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็น และไม่เป็นโรคตับ โดยศึกษาเปรียบเทียบเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก เกณฑ์ในการเปรียบเทียบคือ ค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate: APER) โดยวิธีที่มีประสิทธิภาพสูงสุดคือ วิธีที่มีค่าอัตราการพยากรณ์ที่ผิดพลาดต่ำสุด ซึ่งตัวแปรอิสระที่ศึกษาในงานวิจัยนี้มี 8 ตัว คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin(DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) ซึ่งผลการศึกษาสามารถสรุปได้ ดังนี้

ผลการวิเคราะห์จำแนกกลุ่ม พบว่า ฟังก์ชั่นที่ใช้ในการจำแนกกลุ่มคือ

$$\hat{Y}_1 = -109.090+12.996x_1+11.447x_2+12.969x_3+16.128x_4-4.460x_5+0.238x_6-0.015x_7+0.245x_8 +105.718x_1x_2+105.896x_1x_3+0.158x_1x_4-1.152x_1x_5+0.114x_1x_6-0.052x_1x_7-0.018x_1x_8-103.452x_2x_3-0.350x_2x_4 -3.432x_2x_5-0.154x_2x_6+0.102x_2x_7-0.002x_2x_8-1.702x_3x_4+6.850x_3x_5-0.050x_3x_6+0.002x_3x_7+0.018x_3x_8 +13.890x_4x_5+0.048x_4x_6+0.018x_4x_7-0.040x_4x_8-0.158\ x_5x_6+0.006x_5x_7+0.126x_5x_8+0.006x_6x_7-53.741x_1^2 -54.977x_2^2-56.769x_3^2-8.797x_4^2-51.671x_5^2-0.008x_6^2-0.005x_7^2-0.001x_8^2$$

$$\hat{Y}_2 = -27.429-1.910x_1+6.803x_2+6.263x_3+1.581x_4-0.866x_5+0.018x_6-0.006x_7+0.006x_8+112.122x_1x_2 +112.296x_1x_3+0.608x_1x_4-0.778x_1x_5+0.012x_1x_6-0.002x_1x_7+0.002x_1x_8-112.030x_2x_3-0.306x_2x_4+0.060x_2x_5 -0.014x_2x_6+0.006x_2x_7-0.002x_2x_8-0.484x_3x_4+0.612x_3x_5-0.006x_3x_6-0.002x_3x_7+0.002x_3x_8+3.392x_4x_5-0.006x_4x_6 +0.002x_4x_7-0.004x_4x_8+0.014x_5x_6-0.004x_5x_7+0.006x_5x_8-56.158x_1^2-56.767x_2^2-56.947x_3^2-1.413x_4^2-2.460x_5^2$$

โดยมีความถูกต้องแม่นยำในการพยากรณ์ 97.67% หรือมีความคลาดเคลื่อนในการพยากรณ์ประมาณ 2.3%

ผลการวิเคราะห์การถดถอยแบบโลจิสติก พบว่า สมการถดถอยคือ

$$\hat{g}(x) = 13.2738 +0.1503(x_1+x_2+x_3) - 3.7401x_4+0.9760x_5 -0.1154x_6+0.00787x_7-0.0466x_8$$

ซึ่งเมื่อทดสอบความเหมาะสมของตัวแบบแล้วพบว่า ไม่เหมาะสมกับข้อมูล แม้ว่าจะให้ความถูกต้องแม่นยำในการพยากรณ์ 94.33% หรือมีค่าความคลาดเคลื่อนในการพยากรณ์ประมาณ 5.7% ก็ตาม

ผลการเปรียบเทียบอัตราความผิดพลาดในการพยากรณ์การจำแนกกลุ่มผู้ป่วยโรคตับ พบว่า การวิเคราะห์จำแนกกลุ่มมีประสิทธิภาพสูงกว่าวิธีการถดถอยโลจิสติก และเนื่องจากการวิเคราะห์การถดถอยโลจิสติกนั้น ตัวแบบที่ได้ไม่เหมาะสมกับลักษณะข้อมูล จึงควรใช้วิธีจำแนกกลุ่มสำหรับข้อมูลชุดนี้

_____ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ

Syamol Lumlongrut 2007: A Comparative Study of Liver Disease Diagnosis Using Discriminant Analysis and Logistic Regression Analysis for the Patients in Chachoengsao Hospital. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Boonorm Chomtee, Ph.D. 91 pages.

The purpose of this research was to compare the two statistical methods : discriminant analysis and logistic regression analysis in liver diagnosis for the patients in Chachoengsao hospital. These patients who were taken the liver function test will be classified to 2 groups : a group of patients who had possibility to get liver disease and a group of patients who had not possibility to got liver disease using the two statistical methods and the criteria for comparison is the apparent error rate(APER) which the lower APER is the higher efficiency method. The liver function test data are Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin(DB: x_5), SGOT (x_6), SGPT (x_7) and Alkaline Phosphatase (ALP: x_8). The results of the two statistical methods are as follows :

For discriminant analysis, the quadratic function is

$\hat{Y}_1 = -109.090+12.996x_1+11.447x_2+12.969x_3+16.128x_4-4.460x_5+0.238x_6-0.015x_7+0.245x_8$
$+105.718x_1x_2+105.896x_1x_3+0.158x_1x_4-1.152x_1x_5+0.114x_1x_6-0.052x_1x_7-0.018x_1x_8-103.452x_2x_3-0.350x_2x_4$
$-3.432x_2x_5-0.154x_2x_6+0.102x_2x_7-0.002x_2x_8-1.702x_3x_4+6.850x_3x_5-0.050x_3x_6+0.002x_3x_7+0.018x_3x_8$
$+13.890x_4x_5+0.048x_4x_6+0.018x_4x_7-0.040x_4x_8-0.158\ x_5x_6+0.006x_5x_7+0.126x_5x_8+0.006x_6x_7-53.741x_1^2$
$-54.977x_2^2-56.769x_3^2-8.797x_4^2-51.671x_5^2-0.008x_6^2-0.005x_7^2-0.001x_8^2.$

$\hat{Y}_2 = -27.429-1.910x_1+6.803x_2+6.263x_3+1.581x_4-0.866x_5+0.018x_6-0.006x_7+0.006x_8+112.122x_1x_2$
$+112.296x_1x_3+0.608x_1x_4-0.778x_1x_5+0.012x_1x_6-0.002x_1x_7+0.002x_1x_8-112.030x_2x_3-0.306x_2x_4+0.060x_2x_5$
$-0.014x_2x_6+0.006x_2x_7-0.002x_2x_8-0.484x_3x_4+0.612x_3x_5-0.006x_3x_6-0.002x_3x_7+0.002x_3x_8+3.392x_4x_5-0.006x_4x_6$
$+0.002x_4x_7-0.004x_4x_8+0.014x_5x_6-0.004x_5x_7+0.006x_5x_8-56.158x_1^2-56.767x_2^2-56.947x_3^2-1.413x_4^2-2.460x_5^2.$

The approximate accuracy of the function is 97.67%.

For logistic regression analysis, the regression model is

$\hat{g}(x) = 13.2738 +0.1503(x_1+ x_2+x_3) - 3.7401x_4+0.9760x_5 -0.1154x_6+0.00787x_7-0.0466x_8.$

However , after testing the logistic regression analysis, it was found that regression model is not appropriate at $\alpha = 0.05$.

For the comparison between these two statistical methods, it is shown that discriminant analysis is more efficiency than logistic regression analysis based on APER. Moreover, discriminant analysis was recommended because of the inappropriate of the logistic regression model.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ผศ.ดร. บุญอ้อม โฉมที ประธานกรรมการที่ปรึกษาวิทยานิพนธ์ รศ. อนันต์ชัย เชื่อนธรรม กรรมการที่ปรึกษาสาขาวิชาเอก และผศ. สาโรจน์ โอพิทักษ์ชีวิน กรรมการที่ปรึกษาสาขาวิชารอง ที่ได้ให้คำปรึกษาในการเรียน การค้นคว้า ตลอดจนการตรวจแก้ไขวิทยานิพนธ์จนกระทั่งเสร็จสมบูรณ์ และกราบขอบพระคุณ ดร. ชนิศวรา เลิศอมรพงษ์ ผู้แทนบัณฑิตวิทยาลัย ที่ได้ให้ความกรุณาตรวจแก้ไขวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณ คุณแม่ ญาติผู้ใหญ่ที่เคารพ รวมทั้งพี่ ๆ เพื่อน ๆ ที่ให้ความช่วยเหลือในหลาย ๆ ด้าน และเป็นกำลังใจในการทำวิทยานิพนธ์จนสำเร็จลุล่วงได้

ด้วยความดี หรือประโยชน์อันใดเนื่องจากวิทยานิพนธ์เล่มนี้ ขอมอบแด่คุณแม่ และคณาจารย์ทุกท่าน ที่ได้อบรมสั่งสอนจนมีความรู้จนถึงปัจจุบัน

ศยามล ลำลองรัตน์

กันยายน 2550

สารบัญ

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

การศึกษาเปรียบเทียบการวินิจฉัยโรคตับด้วยการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก กรณีศึกษาผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา

A Comparative Study of Liver Disease Diagnosis Using Discriminant Analysis and Logistic Regression Analysis for the Patients in Chachoengsao Hospital

คำนำ

มาลินี (2548) ได้กล่าวว่าตับทำหน้าที่สำคัญ และซับซ้อนมากมาย ซึ่งหน้าที่ของตับจะเริ่มตั้งแต่ขณะที่เรายังเป็นทารกอยู่ในครรภ์มารดา โดยตับจะทำหน้าที่สร้างเม็ดเลือดแดง และเมื่อร่างกายเจริญเติบโตขึ้นตับก็จะเปลี่ยนแปลงหน้าที่ โดยมีหน้าที่ที่สำคัญดังนี้

1. หน้าที่เป็นคลังสินค้า

1.1 สร้างสารต่าง ๆ ที่จำเป็นต่อร่างกาย รวมถึงการสร้างไขมัน เช่น คลอเรสเตอรอล และการสร้างน้ำตาลกลูโคสจากสารอื่น ๆ เพื่อนำมาใช้ในร่างกาย

1.2 ผลาญสารอาหารในร่างกายให้เกิดพลังงาน ในทางตรงกันข้าม ขณะที่มีการดูดซึมอาหารตับจะนำกลูโคสไปสร้างเป็นไกลโคเจนเก็บไว้ เพื่อป้องกันกลูโคสในกระแสเลือดไม่ให้เพิ่มขึ้นมาก

1.3 ช่วยควบคุมระดับน้ำตาลในเลือดทำให้สามารถรักษาระดับพลังงานได้สม่ำเสมอตลอดทั้งวัน

1.4 เปลี่ยนอาหารที่เรารับประทานเข้าไปให้เป็นพลังงานที่ถูกสะสมไว้เป็นเคมีธาตุที่จำเป็นสำหรับชีวิต และการเจริญเติบโตของเรา

2. หน้าที่เป็นโรงงานผลิต

2.1 ขับบิลลิรูบิน ซึ่งเป็นสารที่เกิดจากการสลายของเม็ดเลือดแดงที่หมดอายุออกจากร่างกาย

2.2 สร้างน้ำดี และขับออกทางท่อน้ำดีไหลลงสู่ลำไส้เล็ก

2.3 ผลิตโปรตีนที่สำคัญของรางกาย เช่น อัลบูมิน ไฟบริโนเจน โพรทรอมบิน

3. หน้าที่เป็นโรงงานกำจัดของเสีย

3.1 จัดการเกี่ยวกับยาต่าง ๆ ที่ได้รับการดูดซึมมาจากระบบย่อยอาหาร เพื่อให้ร่างกายสามารถได้รับยาต่าง ๆ เหล่านั้นได้อย่างมีประสิทธิภาพสูงสุด และในขณะเดียวกันก็ขจัดส่วนเกินออกไปจากร่างกายด้วย นอกจากนี้ยังทำลายสารพิษต่าง ๆ เช่น แอลกอฮอล์ ยาต่าง ๆ อาทิ ยาปฏิชีวนะ ยากล่อมประสาท และยาอื่น ๆ เพื่อทำให้ยาหมดฤทธิ์ และสามารถขับออกจากร่างกายได้ นอกจากนั้นเซลล์ต่าง ๆ ที่อยู่ในตับจะคอยช่วยตรวจกรองสิ่งแปลกปลอมโดยเฉพาะเชื้อโรคซึ่งมาจากลำไส้ ทำให้ไม่เกิดการติดเชื้อรุนแรง

3.2 เปรียบเสมือนเครื่องกรองที่ขจัดสารพิษต่าง ๆ จากเลือดแล้วแปรพิษต่าง ๆ ให้อยู่ในสภาพของสารที่จะขับออกจากร่างกายได้ เช่น การสลายแอมโมเนียซึ่งเป็นสารที่เกิดจากการสลายกรดอมิโน และโปรตีน หากมีการคั่งของแอมโมเนียในปริมาณมากจะทำให้เรามีความรู้สึกสับสนหรือหมดสติได้ แอมโมเนียจะได้รับการย่อยสลายโดยตับซึ่งจะทำให้กลายเป็นยูเรีย ซึ่งเป็นสารที่มีพิษน้อยกว่า และสามารถกำจัดออกจากร่างกายผ่านทางไตได้ ดังนั้นถ้าตับเสื่อมร่างกายก็จะเต็มไปด้วยสารพิษตกค้าง และในที่สุดจะไม่สามารถมีชีวิตอยู่ได้

การตรวจว่าตับมีการทำงานปกติหรือไม่ แพทย์จะสั่งตรวจการทำงานตับในกรณีตรวจร่างกายประจำปี หรือตรวจในกรณีที่ผู้ป่วยมีอาการตัวเหลือง หรือบวม สำหรับการตรวจการทำงานของตับ (Liver Function Test) ที่ทางโรงพยาบาลเมืองฉะเชิงเทราใช้ในการตรวจผู้ป่วยโดยหลักคือ

1. Total protein (TP): โปรตีนนี้สร้างในตับ ซึ่งมีหลายโรคที่ทำให้ค่าโปรตีนนี้สูง หรือต่ำ ได้แก่ โรคตับ, โรคไต, โรคมะเร็ง และโรคขาดอาหารโปรตีน เป็นต้น มีค่าปกติอยู่ในช่วง 6 – 8 g%

2. Albumin (Alb): เป็นโปรตีนที่สร้างจากตับ หากตับเกิดเสียหายก็ไม่สามารถสร้างโปรตีนชนิดนี้ ทำให้โปรตีนนี้มีระดับต่ำเป็นผลให้เกิดอาการบวมที่เท้า และทั่วร่างกาย โดยมีค่าปกติอยู่ในช่วง 3.4 – 5.5 g%

3. Globulin (Glb): เป็นโปรตีนกลุ่มหนึ่งนอกเหนือจาก Albumin ค่านี้ได้จากการนำค่า Total Protein ลบด้วยค่า Albumin ค่าสูงจะพบในภาวะอักเสบติดเชื้อเรื้อรัง โรคตับ โรคข้ออักเสบ Rheumatoid arthritis และ Lupus ค่าต่ำพบได้ในภาวะโรคไตบางชนิด โรคภูมิคุ้มกันบกพร่อง อื่นๆ โดยมีค่าปกติอยู่ในช่วง 2 – 3.5 g%

4. Total Bilirubin (TB): เป็นน้ำดีชนิดหนึ่งซึ่งหากมีค่าสูงมักเกิดจากโรคตับ และมีการแตกของเม็ดเลือดแดง หรือภาวะฮีโมไลซิส (Hemolysis) โดยมีค่าปกติอยู่ในช่วง 0.1 – 1.1 mg%

5. Direct Bilirubin (DB): เป็นน้ำดีชนิดหนึ่ง ซึ่งในผู้ป่วยโรคนิ่วในถุงน้ำดี และโรคตับจะพบว่ามีค่าสูง ไดเร็คบิลบิรูบิน โดยมีค่าปกติอยู่ในช่วง 0 – 0.5 mg%

6. SGOT: เป็นเอนไซน์ในตับที่ช่วยในการสร้างกรดอะมิโน และโปรตีน หากมีการทำลายหรือการอักเสบของตับจะมีการหลั่งเอนไซน์บางชนิดออกจากตับสู่กระแสเลือด และเป็นเอ็นไซม์ที่พบได้ใน หัวใจ, ตับ และกล้ามเนื้อ ค่าสูงพบในภายหลังภาวะHeart Attack, โรคตับ, โรคการบาดเจ็บของกล้ามเนื้อ โดยมีค่าปกติอยู่ในช่วง 0 – 45 U/L

7. SGPT: เป็นเอนไซน์ในตับที่ช่วยในการสร้างกรดอะมิโน และโปรตีน หากมีการทำลายหรือการอักเสบของตับจะมีการหลั่งเอนไซน์บางชนิดออกจากตับสู่กระแสเลือด และเป็นเอ็นไซม์ที่พบได้ในตับเป็นหลัก ส่วนในกล้ามเนื้อพบได้บ้าง ดังนั้นค่านี้จึงจำเพาะเจาะจงถึงโรคตับมากกว่า SGOT ซึ่งค่าสูงพบในโรคตับ โดยมีค่าปกติอยู่ในช่วง 9 – 72 U/L

8. Alkaline Phosphatase (ALP): เป็นสารเอ็นไซม์ที่สร้างจากตับ และกระดูก โดยค่าสูงขึ้นเป็นตัวบ่งชี้ถึงภาวะโรคของตับ หรือกระดูก และสามารถใช้เป็นตัวบ่งชี้ถึงมะเร็งได้ ค่าที่ต่ำกว่าปกติเป็นตัวบ่งชี้ถึงภาวะขาดโปรตีน ขาดอาหาร ขาดวิตามิน โดยมีค่าปกติอยู่ในช่วง 38 – 126 U/L

ในปัจจุบันการตรวจการทำงานของตับของโรงพยาบาลเมืองฉะเชิงเทรามีการตรวจที่เพิ่มมากขึ้นในทุก ๆ ปี ดังนั้นผู้วิจัยจึงให้ความสนใจในการนำเทคนิคทางสถิติที่สามารถใช้ในการจำแนกกลุ่มผู้ป่วยโรคตับมาทำการศึกษาเพื่อช่วยให้แพทย์สามารถวินิจฉัยโรคตับได้รวดเร็วขึ้น และสามารถให้การรักษาแก่ผู้ป่วยได้ทันการณ์ โดยกัลยา (2548) กล่าวว่า เทคนิคทางสถิติที่ใช้ในการจำแนกกลุ่มมีหลายวิธี ได้แก่ การวิเคราะห์กลุ่ม (Cluster Analysis), การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) และการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นต้น

วิธีการวิเคราะห์กลุ่มเป็นวิธีที่ใช้ในการแบ่งกลุ่มของคน, สัตว์ หรือสิ่งของ ฯลฯ ออกเป็นกลุ่มย่อย โดยให้หน่วยที่อยู่ในกลุ่มเดียวกันมีความคล้ายกันในเรื่องที่จะศึกษา ส่วนหน่วยที่อยู่ต่างกลุ่มกันจะมีความแตกต่างกันในเรื่องที่สนใจศึกษา ซึ่งวิธีการวิเคราะห์กลุ่มนั้นไม่จำเป็นต้องทราบจำนวนกลุ่มย่อยมาก่อนว่ามีกี่กลุ่ม และไม่ทราบมาก่อนว่าหน่วยใดจะอยู่หน่วยใดมาก่อน แต่วิธีการวิเคราะห์นี้จะช่วยพิจารณา และจัดให้ว่าหน่วยใดควรอยู่ในกลุ่มใด โดยศึกษาจากตัวแปรที่คาดว่าจะทำให้หน่วยที่อยู่ต่างกันมีความแตกต่างกัน วิธีการวิเคราะห์นี้จึงไม่มีการแบ่งตัวแปรว่าตัวแปรใดเป็นตัวแปรอิสระ หรือตัวแปรใดเป็นตัวแปรตาม

ส่วนการวิเคราะห์จำแนกกลุ่มเป็นการแบ่งกลุ่มของคน, สัตว์ หรือสิ่งของ ฯลฯ ออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม ซึ่งวิธีการวิเคราะห์นี้จะต้องทราบมาก่อนว่าหน่วยใดจะอยู่หน่วยใด โดยมีวัตถุประสงค์เพื่อหาสาเหตุว่ามีตัวแปร หรือปัจจัยใดบ้างที่ทำให้หน่วยต่างกัน และปัจจัยใดบ้างที่มีความสำคัญต่อหน่วยที่ทำการศึกษา โดยเขียนสมการความสัมพันธ์ระหว่างตัวแปรต้น และตัวแปรอิสระให้อยู่ในรูปเชิงเส้น นอกจากนั้นยังสามารถใช้สมการความสัมพันธ์เชิงเส้นดังกล่าวพยากรณ์กลุ่มให้กับหน่วยใหม่

การวิเคราะห์การถดถอยโลจิสติกมีวัตถุประสงค์เหมือนกับการวิเคราะห์จำแนกกลุ่ม คือ หาปัจจัยที่ทำให้กลุ่มต่างกัน และพยากรณ์กลุ่มให้กับหน่วยใหม่ ซึ่งต้องทราบมาก่อนว่ามีกี่กลุ่ม และหน่วยใดอยู่ในกลุ่มใดมาก่อนเช่นกัน แต่ส่วนที่แตกต่างจากการวิเคราะห์จำแนกกลุ่ม คือ สมการที่แสดงความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระไม่ได้อยู่ในรูปเชิงเส้น และสิ่งที่พยากรณ์คือ จะพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มที่สนใจ

ซึ่งในการวิจัยนี้จะทำการศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติกว่าให้ผลการศึกษาที่เหมือนหรือแตกต่างกัน เนื่องจากทั้ง 2 วิธีนี้สามารถนำไปใช้ในการพยากรณ์กลุ่มให้แก่หน่วยใหม่ได้ โดยพิจารณาจากการจำแนกกลุ่มของข้อมูลว่าวิธีการใดให้ค่าอัตราการพยากรณ์ที่ผิดพลาดต่ำสุด ซึ่งตัวแปรอิสระที่ทำการศึกษาในงานวิจัยนี้มี 8 ตัว คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8)

วัตถุประสงค์

1. ในการวิจัยครั้งนี้ ผู้วิจัยมีวัตถุประสงค์เพื่อศึกษาการวินิจฉัยการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา ที่ได้รับการตรวจการทำงานของตับ (Liver Function Test) โดยพิจารณาจำแนกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และกลุ่มผู้ป่วยที่ไม่มีแนวโน้มเป็นโรคตับ โดยศึกษาเปรียบเทียบเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก โดยพิจารณาจากค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate: APER) ซึ่งวิธีที่มีประสิทธิภาพสูงกว่าเป็นวิธีที่มีค่าอัตราการพยากรณ์ที่ผิดพลาดต่ำกว่า

2. เพื่อให้ได้ตัวแบบ หรือสมการที่ใช้ในการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย

ขอบเขตของการวิจัย

1. ผู้ป่วยที่เป็นกลุ่มตัวอย่างในการศึกษาวิจัยครั้งนี้เป็นผู้ป่วยในของโรงพยาบาลเมืองฉะเชิงเทราที่ได้รับการตรวจการทำงานของตับระหว่างเดือนกันยายน 2548 – ตุลาคม 2549 ซึ่งมีจำนวนทั้งสิ้นประมาณ 1200 ราย และกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทราได้ทำการจำแนกกลุ่มผู้ป่วยโรคตับออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และกลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ ซึ่งผู้วิจัยได้สุ่มตัวอย่างผู้ป่วยทั้ง 2 กลุ่มมาศึกษากลุ่มละ 150 ราย

2. ข้อมูลผลการตรวจการทำงานของตับของผู้ป่วยโรงพยาบาลเมือง ฉะเชิงเทราที่นำมาศึกษานี้เป็นข้อมูลผลการตรวจครั้งแรกของผู้ป่วย

3. การจำแนกกลุ่มของผู้ป่วยจำนวน 300 รายที่นำมาศึกษาในงานวิจัยนี้ ออกเป็น 2 กลุ่ม คือ ผู้ป่วยที่มีแนวโน้มเป็นโรคตับจำนวน 150 ราย และผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับจำนวน 150 ราย นั้น กลุ่มงานเทคนิคการแพทย์ โรงพยาบาลเมืองฉะเชิงเทราเป็นผู้ดำเนินการแบ่งกลุ่มให้ในเบื้องต้น โดยพิจารณาจากค่า Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) ซึ่งจะเป็นตัวแปรอิสระจำนวน 8 ตัวแปรในการศึกษาครั้งนี้

4. ในการวิจัยนี้ผู้วิจัยใช้โปรแกรม Statistical Analysis System (SAS) Version 9.1 ในการวิเคราะห์ข้อมูล

5. ในการทดสอบสมมติฐานต่าง ๆ ในงานวิจัยนี้ จะพิจารณาที่ระดับนัยสำคัญ 0.05

ประโยชน์ที่คาดว่าจะได้รับ

1. คาดว่าได้ตัวแบบสำหรับการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย

2. จากการศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก ทำให้ทราบถึงความสามารถในการจำแนกของทั้ง 2 เทคนิคว่ามีความเหมือน หรือแตกต่างกันหรือไม่ อย่างไร

3. สามารถนำผลงานวิจัยนี้ไปใช้ในการจำแนกกลุ่มของผู้ป่วยที่ได้รับการตรวจการทำงานของตับว่าผู้ป่วยมีแนวโน้มเป็น หรือไม่เป็นโรคตับ เพื่อช่วยให้แพทย์สามารถวินิจฉัยโรคตับได้รวดเร็วขึ้น และสามารถให้การรักษาแก่ผู้ป่วยได้ทันการณ์

การตรวจเอกสาร

การตรวจสอบเอกสารแบ่งออกเป็น 2 ส่วน คือ ส่วนของวิธีการทางสถิติที่ใช้ในการวิจัย และส่วนของงานวิจัยที่เกี่ยวข้อง

วิธีการทางสถิติ

1. การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis: DA)

กัลยา (2548) และสุชาต (2548) กล่าวว่า การวิเคราะห์จำแนกกลุ่มเป็นเทคนิคการวิเคราะห์ตัวแปรพหุ (Multivariate Analysis) ที่มีวัตถุประสงค์เพื่อสร้างสมการ หรือตัวแบบขึ้นมาใหม่จากตัวแปรอิสระหลายตัว เพื่อตัวแปรใหม่สามารถจำแนกกลุ่ม หรือระดับของตัวแปรตามให้ได้มากที่สุด โดยที่ตัวแปรตามเป็นตัวแปรเชิงกลุ่ม (Categorical Variable) และตัวแปรอิสระอาจเป็นตัวแปรอิสระอาจเป็นตัวแปรเชิงกลุ่ม หรือตัวแปรต่อเนื่องก็ได้ ซึ่งการวิเคราะห์จำแนกกลุ่มนี้ถูกคิดค้นโดย R.A Fisher เมื่อปี ค.ศ. 1936 ซึ่งในการวิเคราะห์จำแนกกลุ่มนี้มีข้อสมมติเบื้องต้น (Assumption) ดังนี้

1. ตัวแปรตามเป็นตัวแปรเชิงกลุ่ม โดยมีตั้งแต่ 2 กลุ่มขึ้นไป
2. ขนาดตัวอย่างในแต่ละกลุ่ม $(n_i; i = 1, 2, ..., k)$ มีจำนวนมากกว่า หรือเท่ากับ 2
3. จำนวนตัวแปรจำแนก หรือตัวแปรอิสระ $\left(x_1, x_2, ..., x_p\right)$ ที่ใช้ในการวิเคราะห์ควรมีจำนวนน้อยกว่าขนาดตัวอย่างทั้งหมด (n) ลบสอง หรือ $0 < p < n - 2$
4. ตัวแปรอิสระแต่ละตัวต้องเป็นอิสระกัน
5. ตัวแปรอิสระของตัวอย่างแต่ละกลุ่มถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติพหุ (Multivariate Normal Distribution)
6. เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระแต่ละกลุ่มเท่ากัน

การตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์จำแนกกลุ่ม

1. การตรวจสอบว่าเวกเตอร์ตัวแปรอิสระ **x** มีการแจกแจงแบบปกติพหุหรือไม่

Barbara and Linda (2001) กล่าวว่า เมื่อตัวแปรอิสระที่นำมาศึกษามีทั้งตัวแปรต่อเนื่องและตัวแปรเชิงกลุ่ม จึงเป็นเรื่องยากที่ข้อมูลจะมีการแจกแจงแบบปกติหลายตัวแปร แต่ถ้าจำนวนตัวอย่างที่ใช้ในงานวิจัยมีขนาดใหญ่ และวัตถุประสงค์หลักของงานวิจัยคือ ต้องการสร้างสมการจำแนกประเภทที่มีประสิทธิภาพในการจำแนกกลุ่มข้อมูลให้ถูกต้องมากที่สุด ดังนั้นข้อสมมติดังกล่าวจึงไม่ใช่ข้อสมมติเบื้องต้นที่จำเป็น

2. การตรวจสอบการเท่ากันของเมทริกซ์ความแปรปรวนร่วมสามารถใช้วิธี Box'M ในการตรวจสอบดังนี้

สมมติฐานในการทดสอบ $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ และ $H_1 : \Sigma_i \neq \Sigma_j$

หรือ H_0 : เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระทุกกลุ่มเท่ากัน และ

H_1 : เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระบางคู่ไม่เท่ากัน

สถิติทดสอบ $M = -2\rho \ln\left[\dfrac{n^{\frac{pn}{2}} v}{\prod n_i^{\frac{pn_i}{2}}}\right]$

เมื่อ k = จำนวนกลุ่ม

p = จำนวนตัวแปร

n = ขนาดตัวอย่างทั้งหมด – จำนวนกลุ่ม

n_i = ขนาดตัวอย่างในกลุ่มที่ i - 1

$$v = \frac{\prod \left|\text{Within Sum Square Matrix(i)}\right|^{\frac{n_i}{2}}}{\left|\text{Pooled Sum Square Matrix}\right|^{\frac{n}{2}}}$$

$$\rho = 1 - \left[\sum_{i=1}^{k} \frac{1}{n_i} - \frac{1}{n}\right] \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}$$

$$df = \frac{1}{2}(k-1)p(p+1)$$

จากนั้นนำค่าสถิติทดสอบโดยวิธี Box' M ที่ได้จากการคำนวณมาเปรียบเทียบกับค่าไคกำลังสอง ถ้าค่าสถิติทดสอบโดยวิธี Box' M มีค่ามากกว่าจะปฏิเสธ H_0 หรือกล่าวได้ว่าเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระบางคู่ไม่เท่ากัน

Barbara and Linda (2001) กล่าวว่า หากข้อสมมติเกี่ยวกับความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของกลุ่มที่จะจำแนกนั้น หากข้อมูลไม่สอดคล้องกับข้อสมมติยังสามารถใช้การปรับเกี่ยวกับเมทริกซ์ความแปรปรวนร่วมที่นำมาใช้โดยการเปลี่ยนมาพิจารณาเมทริกซ์ความแปรปรวนแบบแยก (Separate variance – covariance matrices) แทนเมทริกซ์ความแปรปรวนร่วมแบบรวม (Pooled variance – covariance matrices) ได้

วิธีการวิเคราะห์จำแนกกลุ่มมีดังต่อไปนี้

1.1 กรณีการวิเคราะห์จำแนกกลุ่มออกเป็น 2 กลุ่ม

เป็นการแบ่งข้อมูลตัวอย่างขนาด n ออกเป็น 2 กลุ่ม คือ กลุ่มที่ 1 ขนาดตัวอย่าง n_1 และกลุ่มที่ 2 ขนาดตัวอย่าง n_2 โดยที่มีตัวแปรอิสระ p ตัว $\left(x_1, x_2, ..., x_p\right)$ และกำหนดให้

กลุ่ม 1: มีเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}_1$ และเมทริกซ์ค่าแปรปรวนร่วม $\boldsymbol{\Sigma}_1$

กลุ่ม 2: มีเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}_2$ และเมทริกซ์ค่าแปรปรวนร่วม $\boldsymbol{\Sigma}_2$

โดยที่ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{p\times p}$ และ $\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$

จะได้ฟังก์ชันจำแนกประเภทที่อยู่ในรูปเชิงเส้น ดังนี้

$$\hat{\mathbf{Y}} = b_1x_1 + b_2x_2 + ... + b_px_p \qquad (1)$$

หรือ $\mathbf{Y} = \mathbf{b}'\mathbf{x}$

เมื่อ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรอิสระ p ตัว โดย $\mathbf{x}' = \left(x_1, x_2, ..., x_p\right)$

$\mathbf{b}$ เป็นเวกเตอร์สัมประสิทธิ์จำแนกประเภท โดย $\mathbf{b}' = \left(b_1, b_2, ..., b_p\right)$

โดยต้องการให้ค่าเวกเตอร์สัมประสิทธิ์ $\mathbf{b}$ ที่ทำให้อัตราส่วนระหว่างความผันแปรระหว่างกลุ่มกับความผันแปรภายในกลุ่มมีค่ามากที่สุด หรือเป็นการหาค่า $\mathbf{b}$ ที่ทำให้ $L = \dfrac{\text{Between - Groups Sum Square}}{\text{Within Groups Sum Square}}$ มีค่ามากที่สุด

จาก $\mathbf{Y} = \mathbf{b}'\mathbf{x}$

จะได้ว่า ค่าเฉลี่ยของกลุ่มที่ 1 : $\bar{\mathbf{Y}}_1 = \mathbf{b}'\bar{\mathbf{x}}_1$

ค่าเฉลี่ยของกลุ่มที่ 2 : $\bar{\mathbf{Y}}_2 = \mathbf{b}'\bar{\mathbf{x}}_2$

โดย $\mathbf{X}_1$ เป็นเมทริกซ์ขนาด $p \times n_1$ ของข้อมูลในกลุ่มที่ 1

$\mathbf{X}_2$ เป็นเมทริกซ์ขนาด $p \times n_2$ ของข้อมูลในกลุ่มที่ 2

$\bar{\mathbf{x}}_1$ เป็นเวกเตอร์ค่าเฉลี่ยของตัวแปร p ตัว ในกลุ่มที่ 1

$\bar{\mathbf{x}}_2$ เป็นเวกเตอร์ค่าเฉลี่ยของตัวแปร p ตัว ในกลุ่มที่ 2

เมื่อ Within Groups Sum Square (SSW) $= \sum_{i=1}^{n_1}(\mathbf{Y}_{1i} - \bar{\mathbf{Y}}_1)^2 + \sum_{j=1}^{n_2}(\mathbf{Y}_{2j} - \bar{\mathbf{Y}}_2)^2$ (2)

จากนั้นแทน $\mathbf{Y}_i = \mathbf{b}'\mathbf{x}_i$ และ $\bar{\mathbf{Y}}_i = \mathbf{b}'\bar{\mathbf{x}}_i$ ลงในสมการที่ (1) จะได้

$$\text{SSW} = \sum_{i=1}^{n_1}\mathbf{b}'(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'\mathbf{b} + \sum_{j=1}^{n_2}\mathbf{b}'(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'\mathbf{b}$$

$$= \mathbf{b}'\left[\sum_{i=1}^{n_1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'\right]\mathbf{b} \quad (3)$$

จากที่กำหนดให้ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{p\times p}$ จึงได้ว่า SSW แปรผันตรงกับ $\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}$

ส่วน Between - Groups Sum Square (SSB) $= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2$ (4)

เมื่อแทน $\mathbf{Y} = \mathbf{b}'\mathbf{x}$ ลงในสมการที่ (3) จะได้

$$\text{SSB} = \mathbf{b}'\left[n_1(\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})' + n_2(\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})'\right]\mathbf{b} \quad (5)$$

โดยที่ $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

หรือ $\bar{x}_1 - \bar{x} = \dfrac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1 + n_2} = -\dfrac{n_2}{n_1 + n_2}(\mathbf{d})$ (6)

และ $\bar{x}_2 - \bar{x} = \dfrac{n_1(\bar{x}_2 - \bar{x}_1)}{n_1 + n_2} = -\dfrac{n_1}{n_1 + n_2}(\mathbf{d})$ (7)

โดยที่ $\mathbf{d} = \bar{x}_2 - \bar{x}_1$ เป็นเวกเตอร์ของผลต่างของเวกเตอร์ค่าเฉลี่ยของ 2 กลุ่ม

แทนค่า $\bar{x}_1 - \bar{x}$ และ $\bar{x}_2 - \bar{x}$ ในสมการ (4) จะได้

$$SSB = \mathbf{b}'\left[n_1\left(\frac{n_2}{n_1 + n_2}\right)^2 \mathbf{d}\mathbf{d}' + n_2\left(\frac{n_1}{n_1 + n_2}\right)^2 \mathbf{d}\mathbf{d}' \right]\mathbf{b} \quad (8)$$

หรือ SSB แปรผันตรงกับ $\mathbf{b}'\mathbf{d}\mathbf{d}'\mathbf{b}$

ดังนั้น $L = \dfrac{\text{Between - Groups Sum Square}}{\text{Within Groups Sum Square}} = \dfrac{SSB}{SSW} = \dfrac{\mathbf{b}'\mathbf{d}\mathbf{d}'\mathbf{b}}{\mathbf{b}'\sum\mathbf{b}}$

ซึ่งวิธีของ Fisher คือ การหาเวกเตอร์สัมประสิทธิ์จำแนกประเภท $\mathbf{b}$ ที่ทำให้ L มีค่าสูงสุด

$$\frac{\partial L}{\partial b} = \frac{\left[\left(\mathbf{b}'\sum\mathbf{b}\right)\left(2\mathbf{d}\mathbf{d}'\mathbf{b}\right) - \left(\mathbf{b}'\mathbf{d}\mathbf{d}'\mathbf{b}\right)\left(2\sum\mathbf{b}\right)\right]}{\left(\mathbf{b}'\sum\mathbf{b}\right)^2} = 0 \quad (9)$$

นำ $\dfrac{1}{2}\mathbf{b}'\sum\mathbf{b}$ คูณสมการที่ (8) จะได้ $\mathbf{d}\mathbf{d}'\mathbf{b} - L\sum\mathbf{b} = 0$

$$\mathbf{b} = \frac{1}{L}\sum{}^{-1}\mathbf{d}\mathbf{d}'\mathbf{b}$$

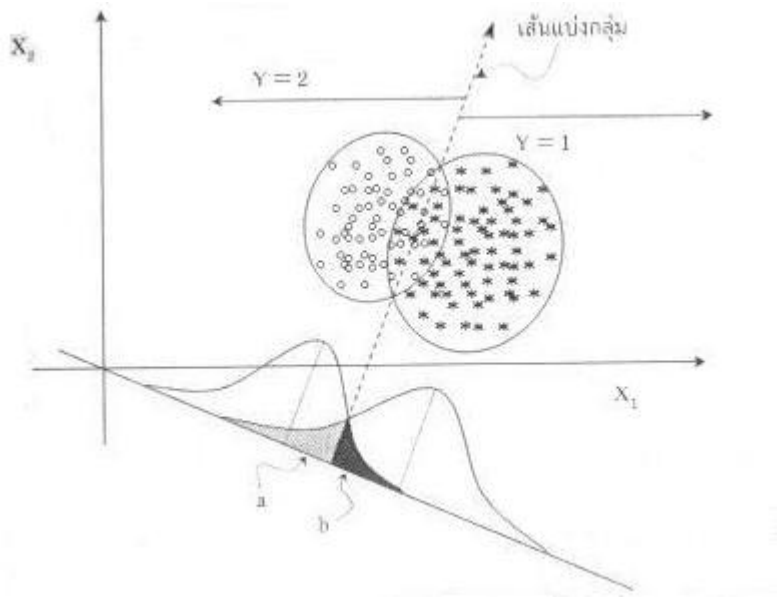
ดังนั้น $\mathbf{b}$ แปรผันตรงกับ $\sum^{-1}\mathbf{d}$

โดยที่ L และ $\mathbf{d}'\mathbf{b}$ เป็นสเกลาร์

สำหรับในกรณีที่ไม่ทราบค่า $\sum$ จะประมาณด้วย $\mathbf{S}_p$ โดยที่ $\mathbf{S}_p$ เป็นเมทริกซ์ค่าแปรปรวนร่วมตัวอย่างแบบรวม (Pooled sample variance – covariance matrix)

ดังนั้น $\mathbf{b} = \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

โดยที่ $\mathbf{S}_p = \frac{1}{n_1 + n_2 - 2}(\bar{\mathbf{x}}_1' \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2' \bar{\mathbf{x}}_2)$



ภาพที่ 1 แสดงการจำแนกโดยใช้ฟังก์ชันการจำแนก
ที่มา: กัลยา (2548)

ภาพที่ 1 เป็นตัวอย่างของการแบ่งกลุ่มออกเป็น 2 กลุ่มโดยมีส่วนเส้นแบ่งกลุ่ม ที่ได้จากสมการ หรือฟังก์ชันจำแนกประเภทในสมการที่ (1) แบ่งกลุ่มตัวอย่างออกเป็น 2 ส่วนในด้านซ้ายและขวา อย่างไรก็ตามเมื่อพิจารณาจากภาพจะเห็นได้ว่ามีกลุ่มตัวอย่างบางรายที่อยู่ในกลุ่มที่ 1 แต่ถูกจัดให้อยู่ในกลุ่มที่ 2 ซึ่งแสดงด้วยพื้นที่แรเงา a โดยที่ a เป็นสัดส่วนหรือร้อยละของกลุ่มตัวอย่างที่ 1 แต่ถูกจัดให้อยู่ในกลุ่มที่ 2 ในทำนองเดียวกันพื้นที่แรเงาส่วน b เป็นสัดส่วนของกลุ่มตัวอย่างที่อยู่ในกลุ่มที่ 2 แต่ถูกจัดให้อยู่ในกลุ่มที่ 1 ซึ่งพื้นที่ส่วนที่แรเงาทั้ง 2 ส่วนนี้ คือ a และ b เป็นสัดส่วนของ

ความผิดพลาดในการจำแนกกลุ่ม ดังนั้นในหลักการของการวิเคราะห์จำแนกประเภท คือ ต้องทำให้เกิดสัดส่วนหรือร้อยละของความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

1.1.1 ทดสอบสมมติฐานของการวิเคราะห์จำแนกประเภท

ในการวิเคราะห์จำแนกประเภทนั้น จะต้องทราบจำนวนกลุ่มก่อนที่จะทำการวิเคราะห์ ซึ่งกลุ่มที่แบ่งนี้อาจจะไม่แตกต่างกันก็ได้ จึงต้องทำการตรวจสอบก่อนว่าทั้ง 2 กลุ่มนี้แตกต่างกันจริงหรือไม่ ก่อนที่จะวิเคราะห์ในขั้นต่อไป ซึ่งอาจมีหลายวิธีในการตรวจสอบได้แก่ Wilk's Lambda, Hotelling T^2, Roy's largest root และ Pillai's Trace เป็นต้น ซึ่งในการวิจัยนี้เลือกใช้วิธี Hotelling T^2 ในการทดสอบ

สมมติฐานในการทดสอบ $H_0 : \mu_1 = \mu_2$ และ $H_1 : \mu_1 \neq \mu_2$
หรือ H_0 : ค่าเฉลี่ยของกลุ่มที่ 1 เท่ากับกลุ่มที่ 2 และ
H_1 : ค่าเฉลี่ยของกลุ่มที่ 1 ไม่เท่ากับกลุ่มที่ 2

สถิติทดสอบ $T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \frac{n_1 n_2}{n_1 + n_2} \mathbf{D}^2$

โดยที่ $F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)\, p} \mathbf{D}^2$

เมื่อ $\mathbf{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
n = ขนาดตัวอย่างของกลุ่มที่ i
p = จำนวนตัวแปรอิสระ
df = $p,\ n_1 + n_2 - p - 1$

จากนั้นนำค่าสถิติทดสอบโดยวิธี Hotelling T^2 ที่ได้จากการคำนวณมาเปรียบเทียบกับค่า F ถ้าค่าสถิติทดสอบโดยวิธี Hotelling T^2 มีค่ามากกว่าจะปฏิเสธ H_0 กล่าวคือค่าเฉลี่ยของกลุ่มที่ 1 ไม่เท่ากับกลุ่มที่ 2 หรือตัวแบบที่สร้างสามารถนำไปใช้ในการจำแนกกลุ่ม

1.1.2 การจัดหน่วยตัวอย่างเข้ากลุ่ม (Allocation of Observations to Group)

การกำหนดกลุ่มให้กับหน่วยใหม่ หรือหน่วยที่ยังไม่ถูกจัดกลุ่มนั้น สามารถทำได้โดยการนำสมการจำแนกกลุ่มที่ได้จากการวิเคราะห์จำแนกกลุ่มมาใช้ในการพยากรณ์หน่วยใหม่ โดยมีเทคนิคการจัดกลุ่มหลายวิธี ดังนี้

ก. Simple Classification Function

วิธีการนี้จะใช้สมการจำแนกกลุ่มของ Fisher มาแทนค่าของตัวแปรอิสระหรือตัวแปรจำแนกกลุ่มของหน่วยใหม่ แล้วคำนวณหาค่าตัวแปรตาม ถ้าค่าตัวแปรตามของกลุ่มใดมีค่ามากที่สุด จะจัดหน่วยให้อยู่ในกลุ่มนั้น ซึ่งวิธีการนี้จัดได้ว่าเป็นวิธีการหนึ่งของวิธีการพาราเมตริก

ข. Probabilities of Group Membership

เป็นการคำนวณหาความน่าจะเป็นหรือโอกาสที่หน่วยใหม่จะอยู่ในกลุ่มที่ i; i = 1,2,…,k แล้วนำมาเปรียบเทียบกัน นั่นคือหา P(หน่วยอยู่ในกลุ่มที่ i) ถ้าพบว่า P(หน่วยอยู่ในกลุ่มที่ 2) มากกว่า P(หน่วยอยู่ในกลุ่มที่ 1) จะกำหนดให้หน่วยนั้นอยู่ในกลุ่มที่ 2 แต่วิธีการนี้มีเงื่อนไขว่าตัวแปรต้องมีการแจกแจงแบบปกติพหุ (Multivariate Normal Distribution) และความน่าจะเป็นที่คำนวณจะเป็นความน่าจะเป็นก่อนหน้า (Posterior Probability)

ค. Generalized Distance Function

เป็นการคำนวณหาระยะห่างจากหน่วยที่ต้องการจัดไปยังค่ากลางของกลุ่ม (Group Centroid) ถ้าระยะห่างดังกล่าวห่างจากค่ากลางของกลุ่มใดต่ำสุด จะจัดให้หน่วยดังกล่าวอยู่ในกลุ่มที่มีระยะห่างต่ำสุดนั้น โดยที่การคำนวณหาระยะห่างนั้นจะใช้ Mahalanobis Distance นั่นคือหาค่า Mahalanobis Distance ของหน่วยนั้นกับจุดกลางของกลุ่ม ต่าง ๆ และจะจัดให้หน่วยใหม่อยู่ในกลุ่มที่มีระยะห่างจากหน่วยนั้นถึงจุดกลางกลุ่มต่ำสุด ซึ่งวิธีการนี้จัดได้ว่าเป็นวิธีการหนึ่งของวิธีการนอนพาราเมตริก

สำหรับในงานวิจัยนี้จะนำเทคนิค Simple Classification function มาใช้ในการจัดกลุ่มของหน่วยตัวอย่างใหม่ เนื่องจากงานวิจัยนี้มีการละการทดสอบข้อสมมติเบื้องต้นที่เกี่ยวกับข้อมูลต้องมีการแจกแจงแบบปกติพหุ ซึ่งเทคนิคนี้มีรายละเอียดดังต่อไปนี้

ก. สร้างฟังก์ชันการจำแนกกลุ่ม (Classification Function) ซึ่งเป็นผลรวมเชิงเส้นของตัวแปรอิสระในแต่ละกลุ่ม โดยจำนวนของฟังก์ชันการจำแนกกลุ่มจะมีจำนวนเท่ากับจำนวนกลุ่มที่ต้องการจำแนก ซึ่งฟังก์ชันการจำแนกกลุ่มมีรูปแบบ ดังนี้

$$\hat{\mathbf{Y}} = b_1x_1 + b_2x_2 + ... + b_px_p$$

หรือ $\mathbf{Y} = \mathbf{b}'\mathbf{x}$

เมื่อ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรอิสระ p ตัว โดย $\mathbf{x}' = (x_1, x_2, ..., x_p)$

$\mathbf{b}$ เป็นเวกเตอร์สัมประสิทธิ์จำแนกประเภทโดย $\mathbf{b}' = (b_1, b_2, ..., b_p)$

ข. ในการจำแนกกลุ่มให้พิจารณาจากค่าฟังก์ชันการจำแนกกลุ่ม (Classification Function) โดยจัดให้สมาชิกใด ๆ เข้าเป็นสมาชิกของกลุ่มที่มีค่าฟังก์ชันการจำแนกกลุ่มสูงที่สุด

1.1.3 ตัวอย่างการวิเคราะห์จำแนกกลุ่ม

จากตัวอย่างข้อมูลความสนใจต่อเครื่องตัดหญ้าที่มีการทำงานแบบต้องมีผู้ควบคุมและแบบอัตโนมัติดังที่แสดงในตารางที่ 1 ซึ่งมีตัวแปร 2 ตัวแปรคือ

$\mathbf{x_1}$ = รายได้ครอบครัว (หน่วยพันดอลลาร์)

$\mathbf{x_2}$ = ขนาดพื้นที่ (หน่วยพันตารางฟิต)

ซึ่งจากข้อมูลมีผลการคำนวณดังนี้ (Richard and Dean, 2002)

ตารางที่ 1 แสดงข้อมูลรายได้ครอบครัว (พันดอลลาร์) และขนาดพื้นที่ (พันตารางฟุต)

กลุ่มที่ 1 : แบบมีผู้ควบคุม		กลุ่มที่ 2 : แบบอัตโนมัติ	
x_1	x_2	x_1	x_2
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6

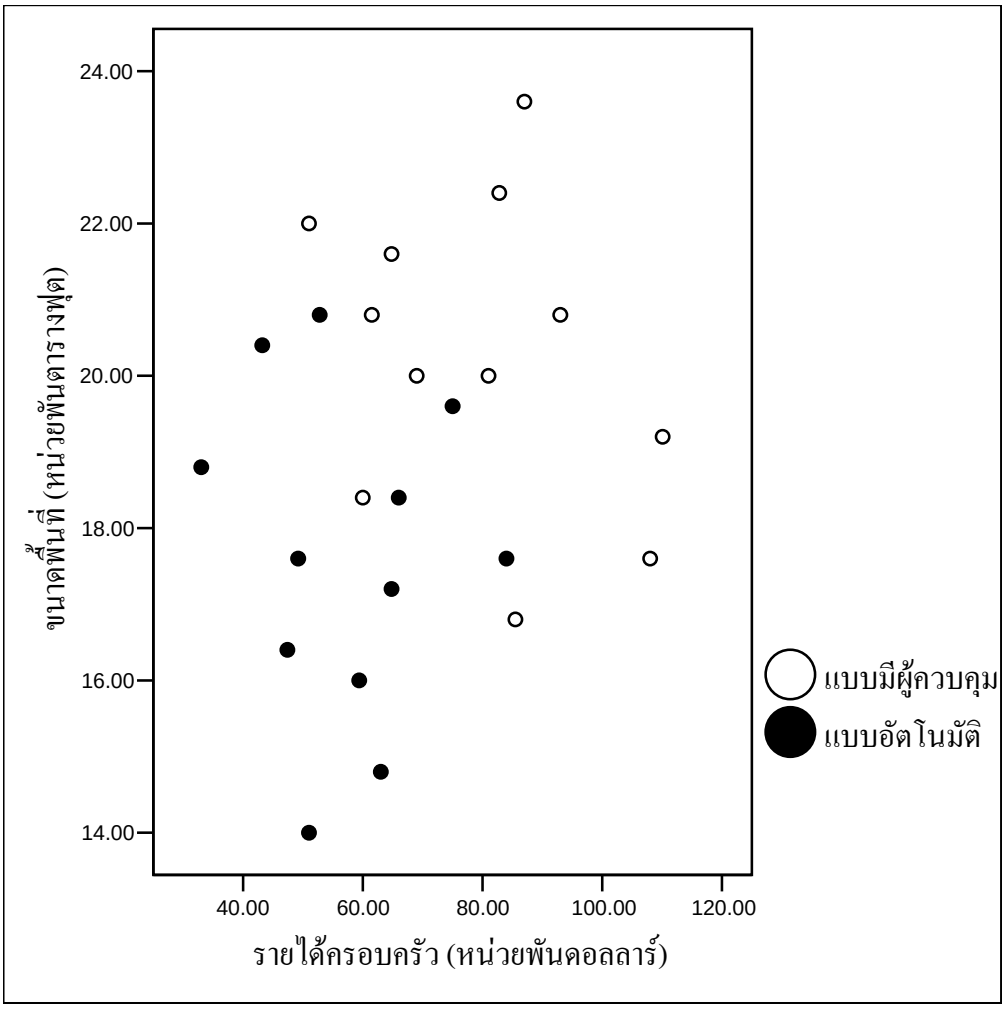
ตารางที่ 1 (ต่อ)

กลุ่มที่ 1 : แบบมีผู้ควบคุม		กลุ่มที่ 2 : แบบอัตโนมัติ	
x_1	x_2	x_1	x_2
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

โดยที่

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 79.475 \\ 20.267 \end{bmatrix} \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} 57.400 \\ 17.633 \end{bmatrix}$$

$$\mathbf{S}_p = \begin{bmatrix} 276.675 & -7.204 \\ -7.204 & 4.273 \end{bmatrix} \qquad \mathbf{S}_p^{-1} = \begin{bmatrix} 0.004 & 0.006 \\ 0.006 & 0.244 \end{bmatrix}$$



ภาพที่ 2 แสดงกราฟระหว่างรายได้ครอบครัว (พันดอลลาร์) และขนาดพื้นที่ (พันตารางฟุต) ต่อความสนใจของเครื่องตัดหญ้าทั้ง 2 แบบ

เมื่อนำข้อมูลจากตารางที่ 1 มาสร้างกราฟจะได้ดังรูปที่ 2 จะเห็นได้ว่ามีกลุ่มตัวอย่างบางรายที่อยู่ในกลุ่มที่ 1 แต่ถูกจัดให้อยู่ในกลุ่มที่ 2 และกลุ่มตัวอย่างที่ 2 บางรายถูกจัดให้อยู่ในกลุ่มที่ 1 แสดงว่าในการจำแนกกลุ่มอาจมีความผิดพลาดเกิดขึ้น

ดังนั้นฟังก์ชันในการจำแนกกลุ่มคือ

$$Y = \mathbf{b}'\mathbf{x} = b_1 x_1 + b_2 x_2$$

$$Y_1 = 0.42959 x_1 + 5.46675 x_2$$

$$Y_2 = 0.32936 x_1 + 4.68157 x_2$$

จากนั้นตรวจสอบสมมติฐานของการวิเคราะห์จำแนกกลุ่มโดยใช้สถิติทดสอบ Hotelling T^2

สมมติฐานในการทดสอบ $H_0 : \mu_1 = \mu_2$ และ $H_1 : \mu_1 \neq \mu_2$
หรือ H_0 : ค่าเฉลี่ยของกลุ่มที่ 1 เท่ากับกลุ่มที่ 2 และ
H_1 : ค่าเฉลี่ยของกลุ่มที่ 1 ไม่เท่ากับกลุ่มที่ 2

ซึ่ง $df = p, n_1 + n_2 - p - 1 = 2, 21$
เมื่อ $\bar{Y}_1 = \mathbf{b}'\bar{\mathbf{x}}_1 = 124.1769$
$\bar{Y}_2 = \mathbf{b}'\bar{\mathbf{x}}_2 = 104.02986$

ดังนั้น $D^2 = \bar{Y}_1 - \bar{Y}_2 = 124.1769 - 104.02986 = 20.14704$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = \frac{(12)(12)}{12+12}(20.14704) = 120.88224$$

โดย $$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 = \frac{12+12-2-1}{2(12+12-2)}(120.88224) = 57.6938$$

เปิดตาราง F ที่องศาอิสระ 2 และ 21 ตามลำดับ ณ ระดับนัยสำคัญ 0.05 จะได้ค่าสถิติ $F_{(2,21)} = 3.47$ ดังนั้นจึงปฏิเสธ H_0 ซึ่งแสดงว่าค่าเฉลี่ยของทั้ง 2 กลุ่มนั้นแตกต่างกันอย่างมีนัยสำคัญทางสถิติ หรือตัวแบบที่สร้างสามารถนำไปใช้ในการจำแนกกลุ่ม

จากตัวอย่างข้างต้น หากต้องการพยากรณ์กลุ่มตัวอย่างของหน่วยใหม่โดยใช้สมการจำแนกกลุ่มที่สร้างไว้ได้ดังนี้

เมื่อ $x_1 = 60$ และ $x_2 = 18$ นำมาแทนค่าในสมการ

$$Y_1 = 0.42959\, x_1 + 5.46675\, x_2 = 124.1769$$
$$Y_2 = 0.32936\, x_1 + 4.68157\, x_2 = 104.02986$$

จากเกณฑ์การจำแนกกลุ่มแบบ Simple Classification function พบว่า Y_1 มีค่ามากกว่า Y_2 แสดงว่าควรจัดหน่วยตัวอย่างใหม่นี้ให้อยู่ในกลุ่มที่ 1 หรือกลุ่มผู้ที่ให้ความสนใจในเครื่องตัดหญ้าแบบมีผู้ควบคุม

1.2 กรณีการวิเคราะห์จำแนกกลุ่มมากกว่า 2 กลุ่ม

ในการจำแนกกลุ่มย่อยที่มากกว่า 2 กลุ่ม หรือแบ่งเป็น k กลุ่มนั้น มีข้อสมมติเบื้องต้นเช่นเดียวกับการจำแนกกลุ่ม 2 กลุ่ม ซึ่งการแบ่งเป็น k กลุ่มนั้นจะสามารถสร้างสมการจำแนกกลุ่มได้เท่ากับ min {p, k – 1} ซึ่งสมการในการจำแนกประเภทมีดังนี้

$$\mathbf{Y_i = b_{i1}x_1 + b_{i2}x_2 + ... + b_{ip}x_p} \; ; i = 1, 2, \ldots, k - 1$$

$\mathbf{Y_i' = b'x_p}$ และหาค่าเฉลี่ย $\mathbf{\bar{Y}_i' = b'\bar{x}_p}$

เมื่อ **x** เป็นเวกเตอร์ของตัวแปรอิสระ p ตัว

b เป็นเวกเตอร์สัมประสิทธิ์จำแนกประเภท

k เป็นจำนวนกลุ่มที่แบ่ง

การวิเคราะห์จำแนกกลุ่มมากกว่า 2 กลุ่ม มีวัตถุประสงค์ดังนี้

ก. ตรวจสอบการแบ่งกลุ่มในรูปแบบกราฟ 2 มิติ และเมื่อเป็นกรณีที่มากกว่า 2 กลุ่มก็ต้องการมากกว่า 1 ฟังก์ชันจำแนก เพื่ออธิบายการแบ่งกลุ่ม ถ้าจุดใน p มิติ Project บนพื้นที่ 2 มิติที่สร้างจากฟังก์ชันจำแนก 2 ฟังก์ชันแรก จะได้จุดสังเกตในการแบ่งกลุ่มนั้น ๆ

ข. หาเซตย่อยของตัวแปรอิสระที่สามารถแบ่งกลุ่มได้ดีเกือบจะเทียบเท่าเซตของตัวแปรอิสระทั้งหมด

ค. จัดลำดับตัวแปรตามความสัมพันธ์ที่เกื้อหนุน (Relative Contribution) กับการแบ่งกลุ่ม

ง. การแปลความหมายของมิติใหม่แสดงได้โดยใช้ฟังก์ชันจำแนก

จ. ใช้ใน Fixed – Effects MANOVA

สมมติข้อมูลมี k กลุ่ม และค่าสังเกต n_i ตัวสำหรับกลุ่มที่ I นั้นจะแปลงแต่ละเวกเตอร์ค่าสังเกต x_{ij} จะได้ $\mathbf{Y_{ij} = b'x_{ij}}$, i = 1, 2, …, k; j = 1, 2, …, n_i และหาค่าเฉลี่ย $\mathbf{\bar{Y}_i' = b'\bar{x}_i}$ เช่นเดียวกับกรณีที่แบ่งออกเป็น 2 กลุ่ม คือ จะหาเวกเตอร์ **b** ที่สามารถจำแนก $\bar{Y}_1, \bar{Y}_2, ..., \bar{Y}_k$ ได้มากที่สุด $L = \dfrac{\mathbf{b'dd'b}}{\mathbf{b'\sum b}}$

สำหรับกรณี k กลุ่ม จะใช้เมทริกซ์ **H** จาก MANOVA แทนที่ $\mathbf{dd'}$ และ **E** แทนที่ Σ จะได้ $L = \dfrac{\mathbf{b'Hb}}{\mathbf{b'Eb}}$ และสามารถเขียนให้อยู่ในรูป $\mathbf{b'Hb} = L\mathbf{b'Eb}$ หรือ $\mathbf{b'}\left(\mathbf{Hb} - L\mathbf{Eb}\right) = 0$

โดยทำให้ L มีค่ามากที่สุด ซึ่งคำตอบที่ว่า $\mathbf{b' = 0'}$ นั้นจะทำให้ L = 0/0 ดังนั้นคำตอบอื่น ๆ สามารถหาได้จาก $\mathbf{Hb} - L\mathbf{Eb} = 0$ ซึ่งสามารถเขียนได้ในรูป $\left(\mathbf{E^{-1}H} - L\mathbf{I}\right)\mathbf{b} = 0$

คำตอบที่ได้ คือ ค่าไอเกน และเวกเตอร์ไอเกนที่สัมพันธ์กันของ $\mathbf{E^{-1}H}$ เมื่อ $L_1 > L_2 > ... > L_s$ และจำนวนของเวกเตอร์ไอเกนที่ไม่เป็นศูนย์ (s) คือ rank ของเมทริกซ์ **H** ซึ่งสามารถหาได้จากค่าที่น้อยกว่าระหว่าง k – 1 และ p โดยค่าไอเกนที่มากที่สุด คือ ค่าที่มากที่สุดของ $L = \dfrac{\mathbf{b'Hb}}{\mathbf{b'Eb}}$ และสัมประสิทธ์ของเวกเตอร์ที่ให้ค่าสูงสุด คือ เวกเตอร์ไอเกนที่สัมพันธ์กัน ดังนั้นฟังก์ชันจำแนกที่ให้จำแนกค่าเฉลี่ยมากที่สุด คือ $\mathbf{Y_1 = b_1x}$

จากเวกเตอร์ไอเกนทั้ง **s** เวกเตอร์ $b_1, b_2, ..., b_s$ ของ $\mathbf{E^{-1}H}$ ซึ่งสัมพันธ์กับ $L_1, L_2, ..., L_s$ จะทำให้ได้ฟังก์ชันจำแนก s ฟังก์ชัน $\mathbf{Y_1 = b_1'x, Y_1 = b_1'x, ..., Y_s = b_s'x}$ ซึ่งไม่สัมพันธ์กัน ฟังก์ชันเหล่านี้จะแสดงมิติ หรือทิศทางของความแตกต่างระหว่าง $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$

ความสำคัญที่สัมพันธ์กันของแต่ละฟังก์ชันจำแนกสามารถกำหนดได้โดยพิจารณาค่าไอเกนที่เป็นสัดส่วนเทียบกับผลรวม ดังนี้ $\dfrac{L_i}{\sum_j L_j}$

จากเกณฑ์นี้ การใช้ฟังก์ชันจำแนก 2 หรือ 3 ฟังก์ชันก็มักจะเพียงพอที่จะอธิบายความแตกต่างระหว่างกลุ่ม ฟังก์ชันจำแนกที่สัมพันธ์กับค่าไอเกนที่เล็ก ๆ ก็จะถูกละทิ้งไป การทดสอบความมีนัยสำคัญของแต่ละฟังก์ชันจำแนกก็มีประโยชน์เช่นกัน

1.3 สถิติที่สำคัญของการวิเคราะห์จำแนกกลุ่ม

ในการนำเทคนิคการวิเคราะห์การจำแนกกลุ่มไปใช้ จะมีสถิติบางตัวที่ผู้ใช้จะต้องทำความเข้าใจ ดังนี้

1.3.1 Eigenvalue คือค่าอัตราส่วนการผันแปรระหว่างกลุ่มต่อการผันแปรภายในกลุ่ม ใช้วัดความสำคัญเชิงเปรียบเทียบของสมการว่า สมการที่ได้มีอำนาจในการแบ่งแยกการเป็นสมาชิกของกลุ่มได้ดีเพียงใด

1.3.2 Relative Percentage หรือCumulative Percentage เป็นค่าที่แสดงความสัมพันธ์ของ Eigenvalue กับสมการที่ได้ว่ามีอำนาจในการแบ่งแยกคิดเป็นร้อยละเท่าใดของอำนาจในการจำแนกรวม

2. การวิเคราะห์การถดถอยแบบโลจิสติก (Logistic Regression Analysis: LA)

การวิเคราะห์การถดถอยแบบโลจิสติกเป็นวิธีการทางสถิติอีกวิธีการหนึ่งที่ถูกใช้ในการศึกษาการพยากรณ์โอกาสที่จะเกิดของเหตุการณ์ ซึ่งการวิเคราะห์การถดถอยแบบโลจิสติกนี้จะมีตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระเป็นได้ทั้งตัวแปรเชิงปริมาณ และตัวแปรเชิงคุณภาพ การวิเคราะห์การถดถอยแบบโลจิสติกนั้นต่างจากการวิเคราะห์จำแนกกลุ่มที่การวิเคราะห์การถดถอยแบบโลจิสติกนั้นไม่มีข้อสมมติเบื้องต้นเกี่ยวกับการแจกแจงของตัวแปรอิสระ และเกี่ยวกับเมทริกซ์ค่าแปรปรวนร่วม ซึ่งการวิเคราะห์การถดถอยแบบโลจิสติกนั้นมี 2 แบบคือ Binary Logistic Regression และ Multinomial Logistic Regression

2.1 Binary Logistic Regression

จะใช้เมื่อตัวแปรตาม Y เป็นตัวแปรเชิงกลุ่มที่มีค่าได้เพียง 2 ค่า คือ 0 กับ 1 ดังนั้นตัวแปรตาม Y จะมีการแจกแจงแบบเบอร์นูลี (Bernoulli distribution) ซึ่งเป็นวิธีการหนึ่งในการพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ จากสมการที่เหมาะสม โดยการเลือกตัวแปรอิสระที่เหมาะสมเพื่อทำให้เปอร์เซ็นต์ของความถูกต้องในการพยากรณ์มีค่าสูงสุด ซึ่งการวิเคราะห์การถดถอยโลจิสติก จะมีเงื่อนไขน้อยกว่าการวิเคราะห์การถดถอยแบบปกติ แต่อย่างไรก็ตาม การวิเคราะห์ความถดถอยโลจิสติกก็ยังมีข้อสมมติเบื้องต้นหลายข้อดังนี้

1. ตัวแปรอิสระอาจจะเป็นข้อมูลที่มีค่าได้ 2 ค่า หรือเป็นสเกลอันตรภาค (Interval Scale) หรือสเกลอัตราส่วน (Ratio Scale) ก็ได้

2. ค่าคาดหวังของค่าคาดเคลื่อนเป็นศูนย์ หรือ $E(e) = 0$

3. e_i และ e_j เป็นอิสระกัน; $i, j = 1, 2, 3, \ldots, n$ (เมื่อ n คือ จำนวนข้อมูลทั้งหมด)

4. e_i และ x_i เป็นอิสระกัน; $i = 1, 2, 3, \ldots, n$ (เมื่อ n คือ จำนวนข้อมูลทั้งหมด)

5. ตัวแปรอิสระไม่ควรมีความสัมพันธ์กัน หรือไม่ควรเกิดปัญหา Multicollinearity ซึ่งการตรวจสอบว่าตัวแปรอิสระมีความสัมพันธ์กันหรือไม่สามารถพิจารณาจากค่า Variance Inflation Factor (VIF) โดยมีหลักเกณฑ์ในการพิจารณาคือ ถ้าค่า VIF มากกว่า 10 แสดงว่าตัวแปรอิสระนั้น ๆ มีความสัมพันธ์กัน

$$\text{โดยที่ } (VIF)_j = \frac{1}{1-R_j^2};\ j = 1, 2, \ldots p$$

เมื่อ R_j^2 คือ ค่าสัมประสิทธิ์ตัวกำหนดจากรูปแบบการถดถอยที่ j

หากข้อมูลตัวแปรอิสระที่นำมาศึกษามีความสัมพันธ์กัน ซึ่งจะทำให้เกิดปัญหา Multicollinearity ซึ่งมีวิธีการแก้ไขปัญหากรณีที่ตัวแปรอิสระที่นำมาศึกษามีความสัมพันธ์กันอยู่หลายวิธีได้แก่ ตัดตัวแปรใดตัวแปรหนึ่งที่สัมพันธ์กันมากออกไป, สร้างตัวแปรใหม่จากตัวแปรเดิมที่มีความสัมพันธ์กัน และการขจัดผลของตัวแปรหนึ่งออกจากอีกตัวแปรหนึ่ง ซึ่งในงานวิจัยนี้เลือกใช้วิธีการสร้างตัวแปรใหม่จากตัวแปรเดิมที่มีความสัมพันธ์กัน

โดยทำให้ค่าของตัวแปรที่มีความสัมพันธ์กันเป็นค่ามาตรฐาน (Standardized Score)

$$z_i = \frac{x_i - \bar{x}}{s}$$

เมื่อ x_i คือ ตัวแปรอิสระที่ i; i = 1, 2, …, n

$\bar{x}$ คือ ค่าเฉลี่ยของตัวแปรอิสระ

s คือ ความแปรปรวน

แล้วจึงนำตัวแปรเหล่านั้นมารวมกันเพื่อสร้างเป็นตัวแปรใหม่

สำหรับ Binary Logistic Regression จะแบ่งออกเป็น 2 กรณี คือ

2.1.1 กรณีที่มีตัวแปรอิสระ 1 ตัว

สมการความถดถอยเชิงเส้นอย่างง่าย หรือสมการที่แสดงความสัมพันธ์ระหว่าง Y กับ x จะอยู่ในรูปแบบเชิงเส้นดังนี้

$$Y = \beta_0 + \beta_1 x_1 + e$$

หรือ $E(Y) = \beta_0 + \beta_1 x_1 + e$ โดยที่ $-\alpha < E(Y) < \alpha$ (10)

สำหรับในการวิเคราะห์ความถดถอยโลจิสติกนั้น เมื่อ Y มีได้เพียง 2 ค่า จะพบว่าความสัมพันธ์ระหว่าง x และY ไม่ได้อยู่ในรูปเชิงเส้น แต่จะอยู่ในรูปดังนี้

$$E(Y) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad \text{โดยที่} \quad 0 \le E(Y) \le 1 \qquad (11)$$

หรือ E(Y) = P(เกิดเหตุการณ์ที่สนใจ) = P(ไม่เกิดเหตุการณ์ที่สนใจ)

2.1.2 กรณีที่มีตัวแปรอิสระมากกว่า 1 ตัว

เมื่อมีตัวแปรอิสระมากกว่า 1 ตัว หรือมีตัวแปรอิสระ p ตัว จะได้ว่า

$$\text{P(เกิดเหตุการณ์)} = \frac{e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}} \qquad (12)$$

$$\text{P(ไม่เกิดเหตุการณ์)} = 1 - \text{P(เกิดเหตุการณ์)}$$

ซึ่งการวิเคราะห์ความถดถอยโลจิสติกนั้นความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระไม่ได้อยู่ในรูปเชิงเส้น จึงมีการปรับให้ความสัมพันธ์อยู่ในรูปเชิงเส้นโดยให้

$$\text{Odds Ratio} = \frac{\text{P(เกิดเหตุการณ์)}}{\text{P(ไม่เกิดเหตุการณ์)}}$$

$$= \frac{P}{1-P}$$

$$= \frac{\dfrac{e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}}}{\dfrac{1}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}}}$$

$$= e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p} \qquad (13)$$

เมื่อ Odds หรือ Odd Ratio แสดงถึงโอกาสที่จะเกิดเหตุการณ์เป็นกี่เท่าของโอกาสที่จะไม่เกิด ถ้าค่า Odd Ratio มากกว่า 1 แสดงว่าโอกาสการเกิดเหตุการณ์มากกว่าการไม่เกิดเหตุการณ์

เมื่อนำสมการที่ (13) มาหาค่า log จะได้

$$\hat{g}(x) = \text{Log(Odds)} = \log(e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}) = \beta_0+\beta_1x_1+\ldots+\beta_px_p \qquad (14)$$

ซึ่งสมการที่ (14) เรียกว่าฟังก์ชันตอบสนองโลจิท (Logit Response Function)

นอกจากนั้นยังมีตัวแบบโพรบิทที่นำมาใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ เมื่อตัวแปรตามเป็นตัวแปรเชิงคุณกลุ่มที่มี 2 ลักษณะ คือ มีค่าเท่ากับ 1 และ 0 ซึ่งการวิเคราะห์ตัวแบบโพรบิทได้ใช้ฟังก์ชันการแจกแจงสะสมปกติมาตรฐาน (Standard normal cumulative distribution function) แปลงค่าเฉลี่ยของตัวแปลตาม (Z_i) ให้เป็นความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (P_i) (ชนิศวรา, 2543) โดยมีตัวแบบดังนี้

$$F^{-1}(P_i) = \boldsymbol{\beta}'\mathbf{x_i} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

ในด้านการคำนวณตัวแบบโพรบิท และตัวแบบโลจิทจะให้ผลใกล้เคียงกัน แต่ตัวแบบโพรบิทมีการคำนวณที่ยุ่งยากกว่า กล่าวคือ ในการกำหนดค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจจะต้องติดอยู่ในรูปแบบของอินทิเกรตเสมอ จึงไม่สะดวก และใช้เวลาในการคำนวณด้วยคอมพิวเตอร์นานกว่า แต่ตัวแบบโลจิทคำนวณได้ง่ายไม่ติดอยู่ในรูปเครื่องหมาย อินทิเกรต ทำให้สามารถประมาณค่าได้ (Judge และคณะ, 1998) ดังนั้นในงานวิจัยนี้จึงเลือกใช้ตัวแบบโลจิทในการวิจัย

2.1.3 การประมาณค่าสัมประสิทธิ์ถดถอยโลจิสติก

ในการประมาณค่าสัมประสิทธิ์การถดถอยมีอยู่หลายวิธีด้วยกัน เช่น วิธีกำลังสองน้อยที่สุด (Least Square Method) และวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) เป็นต้น ซึ่งในงานวิจัยนี้จะใช้วิธีภาวะน่าจะเป็นสูงสุด โดยเริ่มจากการสร้างฟังก์ชันภาวะน่าจะเป็น(Likelihood Function) L ก่อน หลังจากนั้นจึงหาสมการภาวะน่าจะเป็นสูงสุดต่อไป

ถ้ากำหนดให้ตัวแปรตาม Y มีค่า 0 หรือ 1 ดังนั้น P(x) คือ นิพจน์ (Expression) ของค่าสัมประสิทธิ์ที่ต้องการประมาณค่าของความน่าจะเป็นแบบมีเงื่อนไขที่ Y = 1 เมื่อกำหนดค่าตัวแปรอิสระ x มาให้ นั่นคือ P(x) = P(Y = 1| X) ส่วน [1 – P(x)] คือ ความน่าจะเป็นแบบมีเงื่อนไขที่ Y = 0 เมื่อกำหนดค่าตัวแปรอิสระ x มาให้

การประมาณสัมประสิทธิ์โดยวิธีภาวะน่าจะเป็นสูงสุดอาศัยการหาอนุพันธ์ของฟังก์ชันน่าจะเป็น L เทียบกับค่าสัมประสิทธิ์ที่ต้องการประมาณ โดย L มีรูปแบบ ดังนี้

$$L = \prod_{j=1}^{n}\left[P(x_j)\right]^{y_j}\left[1-P(x_j)\right]^{1-y_j} = \prod_{j=1}^{n}\left[\left\{\exp\left(\beta_0+\boldsymbol{\beta}'X_j\right)y_j\right\}/\left\{1+\exp\left(\beta_0+\boldsymbol{\beta}'X_j\right)\right\}\right]$$

หาอนุพันธ์เทียบกับ β_0 และ $\boldsymbol{\beta}$ ซึ่งถ้ามี x เพียง 1 ตัว $\boldsymbol{\beta} = \beta_1$ โดยที่ y = 1 เมื่อ x_j มาจากกลุ่ม 1 และ y = 0 เมื่อ x_j มาจากกลุ่มที่ 2 ส่วน $n = \sum_{i=1}^{2} n$ เมื่อแก้สมการแล้วจะได้ค่าประมาณ $\hat{\beta}_0$ และ $\hat{\boldsymbol{\beta}}$ ซึ่งเรียกว่า ตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator)

2.1.4 ทดสอบนัยสำคัญของค่าสัมประสิทธิ์ถดถอยโลจิสติก

การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์ความถดถอยโลจิสติกของตัวแปรอิสระแต่ละตัวในงานวิจัยนี้ได้นำวิธีการทดสอบวอลด์ (Wald Test) มาใช้ซึ่งเป็นการเปรียบเทียบว่า $\hat{\beta}_i$ กับค่าคลาดเคลื่อนมาตรฐาน (Standard Error) ของค่า $\hat{\beta}_i$ โดยมีวิธีดังนี้

สมมติฐานในการทดสอบ $H_0 : \hat{\beta}_i = 0$; i = 1,2, … , p และ $H_1 : \hat{\beta}_i \neq 0$
หรือ H_0 : ตัวแปรอิสระแต่ละตัวในตัวแบบไม่มีความสัมพันธ์กับตัวแปรตาม
และ H_1 : ตัวแปรอิสระแต่ละตัวในตัวแบบมีความสัมพันธ์กับตัวแปรตาม

สถิติทดสอบ $W_i = \left[\frac{b_i}{SE(b_i)} \right]^2$

เมื่อ b_i = ค่าสัมประสิทธิ์ของแต่ละตัวแปร
$SE(b_i)$ = ค่าคลาดเคลื่อนมาตรฐานของแต่ละตัวแปร
df = 1

จากนั้นนำค่าสถิติทดสอบวอลด์ที่ได้จากการคำนวณมาเปรียบเทียบกับค่าไคกำลังสอง ถ้าค่าสถิติทดสอบวอลด์มีค่ามากกว่าจะปฏิเสธ H_0 หรือกล่าวได้ว่าตัวแปรอิสระแต่ละตัวในตัวแบบมีความสัมพันธ์กับตัวแปรตาม

2.1.5 การตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจิสติก

การตรวจสอบความเหมาะสมของสมการความถดถอยโลจิสติคมีหลายวิธีได้แก่ Likelihood Ratio Chi – Square, Hosmer และ Lemeshow ซึ่งในงานวิจัยนี้ได้ใช้วิธี Hosmer และ Lemeshow ในการทดสอบ ซึ่งมีรายละเอียดดังนี้

สมมติฐานในการทดสอบ H_0 : ตัวแบบเหมาะสมกับข้อมูล และ H_1 : ตัวแบบไม่เหมาะสมกับข้อมูล

$$\text{สถิติทดสอบ} \quad H\text{-}L = \sum_{i=1}^{p} \frac{\left(\sum_j Y_{ij} - \sum_j \hat{P}_{ij}\right)^2}{\left(\sum_j \hat{P}_{ij}\right)\left[\frac{1-\left(\sum_j \hat{P}_{ij}\right)}{n_i}\right]}$$

เมื่อ Y_{ij} = ค่าตัวแปรตามของหน่วยตัวอย่างที่ j ในกลุ่มที่ i ;

$i = 1, 2, \ldots, p$ และ $j = 1, 2, \ldots, n_i$

n_i = จำนวนตัวอย่างของกลุ่มที่ i

$df = p$

$\hat{P}_{ij}$ = ค่าความน่าจะเป็นของหน่วยตัวอย่างที่ j ในกลุ่มที่ i;

$i = 1, 2, \ldots, p$ และ $j = 1, 2, \ldots, n_i$

จากนั้นนำค่าสถิติทดสอบ Hosmer และLemeshow ที่ได้จากการคำนวณมาเปรียบเทียบกับค่าไคคำลังสอง ถ้าค่าสถิติทดสอบ Hosmer และLemeshow มีค่ามากกว่าจะปฏิเสธ H_0 หรือกล่าวได้ว่าตัวแบบไม่เหมาะสมกับข้อมูล

2.1.6 ตัวอย่างการวิเคราะห์ความถดถอยโลจิสติก

จากข้อมูลการศึกษาการซื้อเครื่องสำอางยี่ห้อ A กับรายได้ของลูกค้า(กัลยา, 2548) โดยกำหนดตัวแปร

x_1 แทนรายได้ (หน่วยบาท)

x_2 แทนอายุ (หน่วยปี)

Y แทนตัวแปรตามที่มีค่า 1 (ลูกค้าซื้อเครื่องสำอาง A) และ0 (ลูกค้าไม่ซื้อเครื่องสำอาง A) โดยเก็บข้อมูลของลูกค้าเพศหญิง 60 ราย ซึ่งจากข้อมูลมีผลการคำนวณดังนี้

ตารางที่ 2 ข้อมูลแสดงรายได้(หน่วยบาท) และอายุ(หน่วยปี)

การซื้อ	อายุ (ปี)	รายได้ (บาท)	การซื้อ	อายุ (ปี)	รายได้ (บาท)	การซื้อ	อายุ (ปี)	รายได้ (บาท)
1	40	15,000	0	30	19,000	1	29	32,000
1	45	27,000	0	35	22,000	1	32	12,000
0	20	8,000	1	40	12,000	1	51	51,000
1	30	17,000	0	20	10,000	1	60	20,000
1	55	45,000	0	19	12,000	0	18	10,000
0	15	3,000	0	22	20,000	0	19	12,000
0	18	5,000	0	19	15,000	1	25	9,000
1	20	9,000	1	45	40,000	0	30	9,000
0	25	15,000	1	22	20,000	1	41	30,000
0	40	27,000	1	29	25,000	0	30	20,000
1	30	32,000	1	37	37,000	0	21	18,000
0	27	19,000	1	43	32,000	0	26	31,000
1	35	30,000	1	45	28,000	1	47	43,000
1	42	32,000	1	40	19,000	1	38	32,000
1	53	17,000	0	32	25,000	1	39	29,000
0	35	20,000	1	18	9,000	0	30	35,000
0	18	5,000	1	25	17,000	0	40	41,000
1	27	11,000	1	48	27,000	0	18	17,000
0	19	4,000	1	51	31,000	0	27	20,000
0	30	7,000	1	45	29,000	0	31	30,000

จากการประมวลผลด้วยโปรแกรม SAS ที่ใช้ตัวแบบโลจิทมีผลลัพธ์ดังนี้

ตารางที่ 3 ผลการประมวลการศึกษาการซื้อเครื่องสำอางยี่ห้อ A กับรายได้ของลูกค้า

ตัวแปร	df	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.5680	1.2298	13.7956	0.0002
อายุ	1	0.1626	0.0520	9.7833	0.0018
รายได้	1	-0.00002	0.000042	0.1811	0.6704

เมื่อ $P(\text{เกิดเหตุการณ์}) = \frac{e^{\beta_0+\beta_1x_1+...+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+...+\beta_px_p}}$

ซึ่ง $\text{Odds Ratio} = \frac{P(\text{เกิดเหตุการณ์})}{P(\text{ไม่เกิดเหตุการณ์})} = e^{\beta_0+\beta_1x_1+...+\beta_px_p}$

และ $\hat{g}(x) = \text{Log(Odds)} = \beta_0 + \beta_1x_1 + ... + \beta_px_p$

ดังนั้น $\hat{g}(x) = \text{Log(Odds)} = -4.5680 + 0.1626\,x_1 - 0.00002\,x_2$

ถ้าต้องการพยากรณ์ลูกค้าเพศหญิงที่มีอายุ 32 ปี และมีรายได้ 12,000 บาท จะได้ว่า $\hat{g}(x) = -4.5680 + 0.1626(32) - 0.00002(12000) = 0.3952$

ทำให้ได้ $P(\text{เกิดเหตุการณ์}) = \frac{e^{0.3952}}{1+e^{0.3952}} = 0.5975$

แสดงว่าลูกค้าเพศหญิงที่มีอายุ 32 ปี และมีรายได้ 12,000 บาทมีโอกาสที่จะซื้อเครื่องสำอาง A อยู่ 0.5975 หรือ 59.75%

2.2 Multinomial Logistic Regression

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระหลายตัว กับตัวแปรตาม (Y) ที่เป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 กลุ่ม จะต้องใช้เทคนิค Multinomial Logistic Regression โดยมีเงื่อนไขดังนี้

1. ตัวแปรอิสระบางตัวอาจเป็นตัวแปรเชิงปริมาณ บางตัวอาจเป็นตัวแปรเชิงกลุ่ม
2. ตัวแปรตามต้องเป็นตัวแปรเชิงกลุ่มที่มีมากกว่า 2 ค่า
3. Odds ratio ของตัวแปรเชิงกลุ่ม 2 ค่า จะต้องเป็นอิสระกับค่าอื่น ๆ

ซึ่งคุณสมบัติที่สำคัญของการ Multinomial Logistic Regression คือ เมื่อตัวแปรตามมี 3 กลุ่มย่อยแล้วตัวแบบที่จะใช้จะมี 2 ตัวแบบ ในทำนองเดียวกันถ้าตัวแปรตามมี 4 กลุ่มย่อยแล้ว ตัวแบบที่จะใช้มี 3 ตัวแบบ เนื่องจากอีก 1 กลุ่มจะใช้สำหรับอ้างอิง (Reference Group)

สำหรับงานวิจัยนี้จะนำวิธีการ Binary Logistic Regression มาใช้ในการวิเคราะห์ข้อมูลสำหรับกรณีที่มีตัวแปรอิสระมากกว่า 1 ตัว มาพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับ เพื่อใช้ในการหาสมการการถดถอย และนำไปใช้ในการพยากรณ์

3. การประมาณค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate : APER)

ในงานวิจัยนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการพยากรณ์ผู้ป่วยโรคตับทั้ง 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก ซึ่งสามารถประเมินประสิทธิภาพใน การพยากรณ์ได้โดยการพิจารณาจากอัตราความผิดพลาดในการพยากรณ์ที่เห็นชัด หรือสัดส่วนของผู้ป่วยในกลุ่มตัวอย่างที่พยากรณ์ผิดพลาด ดังนี้

ตารางที่ 4 แสดงผลการจำแนกกลุ่มค่าสังเกตในการจำแนกกลุ่ม

กลุ่มจริง	กลุ่มจากตัวแบบที่สร้าง		ค่าสังเกต
	1	2	
1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

เมื่อ n_{1c} = ขนาดตัวอย่างที่แบ่งกลุ่ม 1 ถูกต้อง

n_{1m} = ขนาดตัวอย่าง 1 จริง แต่ถูกจัดอยู่ในกลุ่ม 2

n_{2m} = ขนาดตัวอย่าง 2 จริง แต่ถูกจัดอยู่ในกลุ่ม 1

n_{2c} = ขนาดตัวอย่างที่แบ่งกลุ่ม 2 ถูกต้อง

ดังนั้น $APER = \frac{n_{1m}+n_{2m}}{n_1+n_2} = \frac{n_{1m}+n_{2m}}{n_{1c}+n_{1m}+n_{2c}+n_{2m}}$

หรืออาจคำนวณอัตราการจัดกลุ่มถูกต้อง (Apparent Correct Classification Rate) ดังนี้

$$\text{Apparent Correct Classification Rate} = \frac{n_{1c}+n_{2c}}{n_1+n_2}$$

หรือ APER = 1 - Apparent Correct Classification Rate

ถ้าอัตราความผิดพลาด (APER) มีค่าน้อย แสดงว่ามีประสิทธิภาพในการพยากรณ์สูง และในการตรงกันข้ามถ้าอัตราความผิดพลาดมีค่ามาก แสดงว่ามีประสิทธิภาพในการพยากรณ์ต่ำ ซึ่งเมื่อคำนวณค่าอัตราความผิดพลาด แล้วจากนั้นผู้วิจัยจะทำการเปรียบเทียบประสิทธิภาพของวิธีการพยากรณ์ 2 วิธีดังกล่าว และสรุปผลการวิจัย

ผลงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่ทำการคิดค้นโดยอรัญญา (2537) เกี่ยวข้องในด้านวิธีการทางสถิติ กล่าวคือ เป็นการเปรียบเทียบการวิเคราะห์การถดถอยแบบโลจิสติก กับการวิเคราะห์จำแนกกลุ่มในการศึกษาปัจจัยที่มีผลต่อความสำเร็จในการศึกษาของนักเรียนนายเรือ โดยจำแนกกลุ่มนักเรียนนายเรือ เป็นกลุ่มประสบความสำเร็จ และไม่ประสบความสำเร็จในการศึกษา ผลการเปรียบเทียบ พบว่า การวิเคราะห์การถดถอยแบบโลจิสติกที่ประมาณพารามิเตอร์ในตัวแบบโดยวิธีภาวะน่าจะเป็นสูงสุด สามารถจำแนกกลุ่มนักเรียนนายเรือ ได้ถูกต้องมากกว่าการวิเคราะห์จำแนกประเภท ทั้งนี้เนื่องจากทั้ง 2 ประชากรไม่มีการแจกแจงแบบปกติ และโควาเรียนซ์ของ 2 ประชากรต่างกัน โดยจำแนกได้ถูกต้องในกลุ่มตัวอย่าง 1 ซึ่งเป็นนักเรียนนายเรือที่ศึกษาชั้นปีที่ 1 ที่โรงเรียนเตรียมทหาร และศึกษาต่อ 4 ปี ที่โรงเรียนเตรียมทหาร คิดเป็นร้อยละ 91.00 และกลุ่มตัวอย่าง 2 เป็นนักเรียนนายเรือที่ศึกษาที่โรงเรียนนายเรือ 5 ปี คิดเป็นร้อยละ 92.73 จากการวิเคราะห์การแบบโลจิสติกของกลุ่มตัวอย่างที่ 1 จะได้ฟังก์ชัน ดังนี้ $\hat{g}(x)$ = -7.1398 + 3.5219 คะแนนเฉลี่ยสะสมจากโรงเรียนเตรียมทหาร + 0.9625 การเข้าร่วมกิจกรรมที่โรงเรียนนายเรือ ส่วนกลุ่มตัวอย่างที่ 2 ได้ฟังก์ชัน ดังนี้ $\hat{g}(x)$ = -7.5163 + 1.6261 คะแนนเฉลี่ยสะสมวิชาวิทยาศาสตร์ + 2.1821 คะแนนเฉลี่ยสะสมวิชาภาษา และสังคมศาสตร์

งานวิจัยที่ทำการคิดค้นโดย ยี่สุ่น (2538) นั้นทำการวิเคราะห์ตัวแปรสำคัญที่สามารถจำแนกกลุ่มนักศึกษาที่เลือก และไม่เลือกเรียนวิชาชีพการพยาบาล เขตกรุงเทพมหานคร โดยมีกลุ่มตัวอย่าง คือ นักเรียนที่เลือกเรียนวิชาชีพการพยาบาล จำนวน 297 คน และนักศึกษาที่ไม่เลือกเรียนวิชาชีพการพยาบาล จำนวน 336 คน จากการสุ่มแบบหลายขั้นตอน เครื่องมือที่ใช้ในการวิจัย คือ แบบสอบถามที่ผู้วิจัยสร้างขึ้น โดยผ่านการตรวจสอบความตรงตามเนื้อหาจากผู้ทรงคุณวุฒิ 10 ท่าน และทดสอบความเที่ยงเท่ากับ 0.90 ข้อมูลที่รวบรวมได้ นำมาวิเคราะห์โดยหา ค่าร้อยละ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และวิเคราะห์ตัวแปรจำแนกประเภท โดยวิธีแบบมีขั้นตอน ผลการวิจัยสรุปได้ว่า ตัวแปรสำคัญที่สามารถจำแนกกลุ่มนักศึกษาที่เลือกเรียนวิชาชีพการพยาบาล และกลุ่มนักศึกษาที่ไม่เลือกเรียนวิชาชีพการพยาบาลมี 11 ตัวแปร เรียงตามลำดับความสำคัญตั้งแต่มากไปหาน้อย คือ การรับรู้ลักษณะวิชาชีพการพยาบาล, อิทธิพลของครอบครัว, รายได้รวมของบิดา และมารดา, ระดับการศึกษาของบิดา, การมีประสบการณ์ดูแลผู้เจ็บป่วย, การรับรู้ความก้าวหน้าในวิชาชีพการพยาบา, อิทธิพลของกลุ่มเพื่อน, อาชีพของบิดาที่เกี่ยวข้องกับสุขภาพ, อิทธิพลของครู – อาจารย์แนะแนว, การรับรู้ความต้องการของตลาดแรงงาน และผลสัมฤทธิ์ทางการเรียนชั้นมัธยมศึกษาตอนปลาย สำหรับสมการจำแนกประเภทที่ได้สามารถใช้คาดประมาณ กลุ่มนักศึกษาที่เลือกเรียนวิชาชีพการพยาบาล และกลุ่มนักศึกษาที่ไม่เลือกเรียนวิชาชีพการพยาบาลได้ถูกต้อง ประมาณร้อยละ 73.78 ได้สมการจำแนกกลุ่มในรูปคะแนนมาตรฐาน ดังนี้ Z = 0.52773 การรับรู้ลักษณะวิชาชีพการพยาบาล +0.47910 อิทธิพลของครอบครัว – 0.35236 รายได้รวมของบิดาและมารดา – 0.24902 ระดับการศึกษาของบิดา + 0.24210 การมีประสบการณ์ดูแลผู้เจ็บป่วย + 0.17118 การรับรู้ความก้าวหน้าในวิชาชีพการพยาบาล – 0.12117 อิทธิพลของกลุ่มเพื่อน + 0.11914 อาชีพของบิดาที่เกี่ยวข้องกับสุขภาพ – 0.10757 อิทธิพลของครู – อาจารย์แนะแนว + 0.10684 การรับรู้ความต้องการของตลาดแรงงาน – 0.07740 ผลสัมฤทธิ์ทางการเรียนชั้นมัธยมศึกษาตอนปลาย

สำหรับงานวิจัยที่ทำการคิดค้นโดยสุวรรณา (2538) นั้นเป็นการสร้างสมการจำแนกกลุ่มการเรียนวิชาชีพประเภทช่างอุตสาหกรรมในระดับประกาศนียบัตรวิชาชีพ อันได้แก่ กลุ่มวิชาชีพช่างอิเล็กทรอนิกส์ ช่างไฟฟ้า ช่างก่อสร้าง ช่างยนต์ และช่างเชื่อม เก็บรวบรวมข้อมูลโดยใช้แบบทดสอบความถนัดทางการเรียนจำนวน 6 ฉบับ ไปทำการทดสอบกับนักเรียนระดับประกาศนียบัตรวิชาชีพชั้นปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2536 ของวิทยาลัยเทคนิค สังกัดกรมอาชีวศึกษาในเขตการศึกษา 8 จำนวน 690 คน วิเคราะห์ข้อมูลด้วยการวิเคราะห์จำแนกกลุ่มแบบขั้นตอนด้วยวิธีการของวิลค์ส โดยมีตัวแปรจำแนก 6 ตัวแปร ได้แก่ ความถนัดทางการเรียนด้านตัวเลข (N) ความถนัดทางการเรียนด้านเชิงจักรกล (M) ความถนัดทางการเรียนด้านการรับรู้ (P) ความถนัดทางการเรียนด้านเหตุผล (R) ความถนัดทางการเรียนด้านมิติสัมพันธ์ (S) ความถนัด

ทางการเรียนด้านภาษา (V) ซึ่งสามารถจำแนกความแตกต่างของกลุ่มนักเรียนทั้ง 5 สาขาวิชาชีพ และได้สมการจำแนกกลุ่ม 4 สมการในรูปคะแนนดิบดังนี้ สมการที่ 1 $\hat{Y}_1$ = 0.0704(N) + 0.0890(M) + 0.0663(P) + 0.0156(R) - 0.0473(s) + 0.1928(V) - 5.0459 สมการที่ 2 $\hat{Y}_2$ = 0.0969 N - 0.0404(M) + 0.1377(P) + 0.0712(R) + 0.0518(S) + 0.0270(V) -2.4516 สมการที่ 3 $\hat{Y}_3$ = 0.0076(N)-0.1105(M) + 0.1752(P) + 0.0330(R) + 0.1177(S) + 0.0464(V) - 3.8479 สมการที่ 4 $\hat{Y}_4$ = 0.1480(N) + 0.1168(M) + 0.0259(P) + 0.1213(R) + 0.0035(s) + 0.0807(V) - 0.1785 ทั้งสี่สมการสามารจำแนกนักเรียนข้าเรียนตามกลุ่มวิชาชีพประเภทช่างอุตสาหกรรมได้อย่างถูกต้อง ร้อยละ 34.78

กนกศรี (2540) ได้ศึกษาการรู้จำลายมือเขียนตัวอักษรภาษาไทยได้นำวิธีการจำแนกกลุ่มทางสถิติมาประยุกต์แก้ปัญหา วิธีการจำนกกลุ่มที่พิจารณามี 3 วิธีคือ การวิเคราะห์จำแนกเชิงเส้น วิธีเพื่อนบ้านใกล้เคียงที่สุด k ค่า โดย k กำหนดค่าเป็น 3, 5 และ 8 และการวิเคราะห์จำแนกประเภทแบบเคอร์เนลที่ใช้ฟังก์ชันปกติเป็นฟังก์ชันหนาแน่นเคอร์เนล ตัวอักษรภาษาไทยที่ยังใช้อยู่ในปัจจุบันจำนวน 42 ตัว จะเขียนโดยใช้ปากกาอิเล็กโทรนิก และแปลงให้เป็นข้อมูลพิกัดจำนวน 5, 8, 12 และ 16 คู่ จากนั้นวิเคราะห์ข้อมูลเพื่อสร้างกฎการจำแนกกลุ่ม การวิเคราะห์จำแนกประเภทแบบเคอร์เนลมีอัตราความผิดพลาดต่ำที่สุด คือ 1.21 เปอร์เซ็นต์ แต่มีข้อเสียในเรื่องความเร็ว และการที่ต้องเก็บตัวอย่างเรียนรู้ทั้งหมดไว้เพื่อใช้ในการจำแนกตัวอักษรภายหลัง การวิเคราะห์จำแนกประเภทแบบเชิงเส้นมีความเร็วสูงกว่า และไม่จำเป็นต้องเก็บตัวอย่างเรียนรู้ แต่อัตราความผิดพลาดสูงกว่า คือ 1.97 เปอร์เซ็นต์ วิธีเพื่อนบ้านใกล้เคียงที่สุด k ค่า มีอัตราการจำแนกกลุ่มผิดพลาดสูงสุด โดยเฉพาะอย่างยิ่งเมื่อ k มีค่ามาก ความแม่นยำในการจำแนกจะเพิ่มขึ้นเมื่อจำนวนตัวอย่างเรียนรู้มากขึ้น และไม่ว่าจะเป็นวิธีการจำแนกกลุ่มแบบใดจำนวนพิกัดที่เหมาะสมคือ 12 คู่

สรินยา (2542) ศึกษาเปรียบเทียบการวิเคราะห์จำแนกกลุ่ม กับการวิเคราะห์การถดถอยมัลติโนเมียล สำหรับการศึกษาปัจจัยที่มีผลต่อระดับคะแนนของนิสิตมหาวิทยาลัยเกษตรศาสตร์ปีเข้าศึกษา 2538 จากการวิเคราะห์ได้นำตัวแปรอิสระทั้ง 9 ตัว ได้แก่ คะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลาย (BEF.ENT), จำนวนหน่วยกิตที่ลงทะเบียนในปีการศึกษา 2541 (CREDIT), ทัศนคติต่อตนเองทางการเรียน (MYSELF5), กลุ่มนิสิต (STUDENT.G), แรงจงใจทางการเรียน (MOTIVE3), คะแนนสอบคัดเลือกเข้ามหาวิทยาลัย (ENT38), พฤติกรรมทางการเรียน (BEHAVE5), แรงจูงใจในการทำงาน (MOTIVE 5) และเพศ (SEX) มาวิเคราะห์ โดยจำแนกนิสิตออกเป็น 4 กลุ่มการวิเคราะห์ คือ กลุ่มนิสิตที่ได้คะแนนเฉลี่ยสะสมต่ำกว่า 2.00 กลุ่มนิสิตที่ได้คะแนนสะสมระหว่าง 2.00 ถึง 2.49 กลุ่มนิสิตที่ได้คะแนนสะสมเฉลี่ยระหว่าง 2.50 ถึง 3.24 และกลุ่มนิสิตที่ได้คะแนนเฉลี่ยสะสม 3.25 ถึง 4.00 พบว่า การวิเคราะห์การถดถอยมัลติโนเมียลที่ประมาณค่าพารามิเตอร์ในตัวแบบโดยวิธี

ภาวะน่าจะเป็นสูงสุดสามารถจำแนกระดับคะแนนเฉลี่ยสะสมของนิสิตมหาวิทยาลัยเกษตรศาสตร์ได้ถูกต้องมากกว่าการวิเคราะห์จำแนก ทั้งนี้เนื่องจากตัวแปรที่ใช้ในการวิเคราะห์บางตัวไม่มีการแจกแจงแบบปกติ และความแปรปรวนร่วมของประชากรต่างกัน โดยจำแนกได้ถูกต้องในกลุ่มนิสิตที่มีระดับคะแนนเฉลี่ยสะสม 2.00 ถึง 2.49 คิดเป็นร้อยละ 73.87 โดยมีสมการการวิเคราะห์ คือ $\hat{Y}_1$ = 16.1616 – 3.1615(BEF.ENT) – 1.3892(BEHAVE6) + 0.4138 (CREDIT) -0.0349(ENT38) + 1.3529(STUDENT.G.) - 1.2940(MOTIVE3) + 0.3822(MOTIVE5) + 0.7006(MYSELF5) – 1.9227(SEX) กลุ่มนิสิตที่มีระดับคะแนนเฉลี่ยสะสม 2.50 ถึง 3.24 คิดเป็นร้อยละ 83.86 โดยสมการการวิเคราะห์ คือ $\hat{Y}_2$ = 3.2479 – 1.4519 (BEF.ENT) – 0.7770(BEHAVE6) + 0.2949(CREDIT) - 0.0158(ENT38) + 1.0223 (STUDENT.G.) – 0.5763(MOTIVE3) – 0.0571(MOTIVE5) + 0.3416(MYSELF5) – 0.6605(SEX) และกลุ่มนิสิตที่มีระดับคะแนนเฉลี่ยสะสม 3.25 ถึง 4.00 คิดเป็นร้อยละ 94.79 โดยมีสมการการวิเคราะห์ คือ $\hat{Y}_3$ = -2.5345 – 0.5120(BEF.ENT) – 0.4006(BEHAVE6) + 0.1668(CREDIT) - 0.0049(ENT38) + 0.0462(STUDENT.G.) – 0.3306(MOTIVE3) - 0.2119(MOTIVE5) + 0.0266(MYSELF5) – 0.1532(SEX)

งานวิจัยที่ศึกษาโดยสุรเชษฐ (2547) เป็นการวิเคราะห์ปัจจัยที่ส่งผลกระทบต่อการชำระคืนหนี้ของเกษตรกร และสร้างสมการจำแนกประเภทเพื่อพยากรณ์เกษตรกรรายใหม่ในการรับเป็นสมาชิกศูนย์พัฒนาว่าเป็นสมาชิกกลุ่มปกติ หรือกลุ่มมีปัญหาในการชำระหนี้คืน จากการศึกษา พบว่ากรณีเกษตรกรมีการชำระคืนหนี้ปกติมีสมการในการจำแนกกลุ่มเพื่อพยากรณ์ ดังนี้ $\hat{Y}_1$ = -13.71 + 0.000082 รายได้สุทธิ + 0.000042 หนี้สินของเกษตรกร + 7.37 ชนิดไม้ดอกที่ปลูก + 7.06 พื้นที่เพาะปลูก ส่วนกรณีเกษตรกรมีปัญหาในการชำระคืนหนี้ปกติมีสมการในการจำแนกประเภทเพื่อพยากรณ์ดังนี้ $\hat{Y}_2$ = -10.92 + 0.000036 รายได้สุทธิ + 0.00015 หนี้สินของเกษตรกร + 5.15 ชนิดไม้ดอกที่ปลูก + 5.77 พื้นที่เพาะปลูก

สำหรับงานวิจัยที่ศึกษาโดยพนิดา (2548) เป็นการวิจัยเพื่อเปรียบเทียบประสิทธิภาพของตัวแบบที่ได้จากการจำแนก และการวิเคราะห์การถดถอยโลจิสติค โดยพิจารณาการจำแนก 3 กรณี คือ (1) จำแนกคนไข้ที่ได้รับการยืนยันการวินิจฉัย เป็นโรคเลปโตสไปโรสิส กับกลุ่มที่วินิจฉัยเป็นโรคอื่น ๆ (2) จำแนกคนไข้ที่ได้รับการยืนยันการวินิจฉัย เป็นโรคเลปโตสไปโรสิส กับกลุ่มที่ที่ได้รับการยืนยันการวินิจฉัย เป็นโรคในกลุ่ม Rickettsiosis (3) จำแนกคนไข้ที่ได้รับการยืนยันการวินิจฉัย เป็นโรคเลปโตสไปโรสิส กลุ่มที่ที่ได้รับการยืนยันการวินิจฉัย เป็นโรคในกลุ่ม Rickettsiosis และกลุ่มที่วินิจฉัยเป็นโรคอื่น ๆ ซึ่งผลการวิจัยโดยรวมพบว่า กรณีที่ 1 การวิเคราะห์การจำแนกมีประสิทธิภาพ

ดีกว่าที่คิดเป็น 79.9% กรณีที่ 2 ทั้ง 2 เทคนิคมีประสิทธิภาพเท่าเทียมกันคิดเป็น 79.0% ส่วนกรณีที่ 3 การวิเคราะห์การถดถอยโลจิสติคมีประสิทธิภาพดีกว่าที่คิดเป็น 62.1%

งานวิจัยที่ทำการคิดค้นโดย Worthington and Grant (1971) ศึกษาปัจจัยที่เกี่ยวข้องกับคะแนนเฉลี่ยสะสมในภาคเรียนแรกของนักเรียนมหาวิทยาลัยยูทาห์ โดยใช้ตัวอย่างนักศึกษาปีที่ 1 ชาย และหญิงจำนวน 1,270 คน และ 990 คน ตามลำดับ โดยวิเคราะห์ความแปรปรวนแบบสองทางและสามทาง พบว่า ปัจจัยที่มีอิทธิพลต่อการศึกษาได้แก่ คะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลาย คะแนนการทดสอบวิชาภาษาอังกฤษ คณิตศาสตร์ สังคมศาสตร์ วิทยาศาสตร์ นโยบายสถาบันการศึกษา เพศ รายได้ครอบครัว จำนวนเด็กในครอบครัว ความเอาใจใส่การเรียนระดับมัธยมศึกษาตอนปลาย

Efron (1975) ศึกษาวิธีการทางสถิติ 2 วิธี ได้แก่ การวิเคราะห์การจำแนกกลุ่ม และการวิเคราะห์การถดถอยแบบโลจิสติก เพื่อใช้ในการจำแนกกลุ่มประชากร โดยที่ประชากร 2 ประชากรมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยต่างกัน และโควาเรียนซ์เมตริกซ์ของ 2 ประชากรเท่ากัน พบว่า เมื่อตัวอย่างมีขนาดใหญ่วิธีการวิเคราะห์จำแนกประเภทจะมีประสิทธิภาพสูงกว่าวิธีการวิเคราะห์การถดถอยแบบโลจิสติก

งานวิจัยที่ทำการศึกษาโดย Press and Wilson (1978) เป็นการศึกษาวิธีการทางสถิติ 2 วิธีการ ได้แก่ การวิเคราะห์การถดถอยแบบโลจิสติก และการวิเคราะห์จำแนกกลุ่ม เพื่อศึกษาปัจจัยที่เกี่ยวข้องกับการเกิดโรคมะเร็งทรวงอกที่สถาบันมะเร็งบริติชโคลัมเบีย ระหว่างปี 1955 - 1963 และศึกษาปัจจัยที่เกี่ยวข้องกับการเปลี่ยนแปลงประชากรใน 50 มลรัฐของสหรัฐอเมริกาในระหว่างปี 1960 – 1970 โดยข้อมูลมีลักษณะผสม คือ มีทั้งข้อมูลเชิงคุณภาพ และข้อมูลเชิงปริมาณ ตัวแปรตามมีค่า 2 ค่า เป็นตัวแปรเชิงคุณภาพ ตัวแปรอิสระมีทั้งตัวแปรเชิงคุณภาพ และตัวแปรเชิงปริมาณ และตัวแปรส่วนใหญ่ไม่มีการแจกแจงแบบปกติ พบว่า การถดถอยแบบโลจิสติกที่ใช้วิธีแมกซิมัมไลลิฮูดประมาณพารามิเตอร์นั้น จะให้ค่า ร้อยละการจำแนกความถูกต้องมากกว่าการวิเคราะห์จำแนกประเภทเล็กน้อย

สำหรับการศึกษาของ Daxin Jiang *et al*. (n.d.) เป็นการศึกษาการจำแนกกลุ่มของข้อมูลการแสดงออกของยีน โดย DNA microarray technology ที่นำมาใช้ในปัจจุบันเป็นการจำลองการแสดงออกของยีนระหว่างกระบวนการทางชีวภาพ และเก็บรวบรวมตัวอย่างของผลต่าง และผลที่สอดคล้อง ความชัดเจนของข้อมูลการแสดงออกของยีนเป็นการยกระดับความรู้ความเข้าใจในเรื่องยีนที่มีกระบวนการทำงานมากมาย และซับซ้อน โดยใช้การวิเคราะห์การจำแนกกลุ่ม เพื่อแสดงให้เห็นธรรมชาติของโครงสร้าง และลักษณะที่น่าสนใจของข้อมูลในเบื้องต้น การวิเคราะห์จำแนกกลุ่มนี้เป็นการสำรวจการแบ่งแยกของกลุ่มข้อมูลบนพื้นฐานของลักษณะพิเศษเฉพาะในแต่ละกลุ่มซึ่งได้มีการพัฒนามานานถึง 30 ปี งานวิจัยนี้เป็นการใช้ microarray technology มาอธิบายบนพื้นฐานปัจจัยของการจัดกลุ่มของข้อมูล การแสดงออกของยีน มาใช้ในการวิเคราะห์การจัดกลุ่มแบบแบ่งแยกสำหรับข้อมูลการแสดงออกของยีนออกเป็น 3 ประเภท

ผลงานวิจัยที่เกี่ยวข้องข้างต้นแสดงตัวอย่างของการสร้างตัวแบบในการพยากรณ์โดยการใช้เทคนิคการวิเคราะห์ทางสถิติ ซึ่งงานวิจัยเหล่านี้เป็นแนวทางในการพิจารณาเลือกใช้การวิเคราะห์ทางสถิติให้เหมาะสมกับวัตถุประสงค์ของงานวิจัย คือ การจำแนกกลุ่มผู้ป่วยโรคตับ และการสร้างตัวแบบในการพยากรณ์หน่วยตัวอย่างใหม่ของผู้ป่วยโรคตับ

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องไมโครคอมพิวเตอร์ที่มีหน่วยความจำขนาด 40 GB มีความเร็วในการประมวลผล 256 MHz ที่ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

2. โปรแกรม Statistical Analysis System (SAS) Version 9.1

วิธีการ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการวินิจฉัยการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา ที่ได้รับการตรวจการทำงานของตับ (Liver Function Test) ซึ่งจะแบ่งกลุ่มผู้ป่วยออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และกลุ่มผู้ป่วยที่ไม่มีแนวโน้มเป็นโรคตับ โดยศึกษาเปรียบเทียบเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) และการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) พร้อมทั้งเปรียบเทียบประสิทธิภาพของการวิเคราะห์ทั้ง 2 วิธีด้วยวิธีการประมาณค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate: APER) และเพื่อให้ได้ตัวแบบที่ใช้ในการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย

ในงานวิจัยนี้ได้สุ่มตัวอย่างผู้ป่วยที่ได้รับการตรวจการทำงานของตับเป็นครั้งแรกในช่วงเดือนกันยายน 3548 – ตุลาคม 2549 ของโรงพยาบาลเมืองฉะเชิงเทรา จำนวน 300 ราย โดยแบ่งออกเป็น 2 กลุ่มคือ ผู้ป่วยที่มีแนวโน้มการเป็นโรคตับ และผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับกลุ่มละ 150 ราย ซึ่งได้รับการอนุเคราะห์จากกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทราเป็นผู้จำแนกผู้ป่วยทั้ง 2 กลุ่มนี้ในเบื้องต้น โดยพิจารณาจากค่าปกติของทั้ง 8 ตัวแปรที่นำมาทำการศึกษาในงานวิจัยนี้ ซึ่งตัวแปรอิสระทั้ง 8 ตัว คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (x_8) โดยขั้นตอนในการศึกษาวิจัยครั้งนี้ แบ่งออกเป็น 3 ขั้นตอนดังนี้

1. การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) ประกอบด้วยขั้นตอนดังนี้

1.1 กำหนดตัวแปร x_i ; i = 1, 2, …, 8 เป็นตัวแปรอิสระที่คาดว่าจะทำให้เกิดความแตกต่างระหว่างกลุ่ม ซึ่งตัวแปรอิสระที่ใช้ในการวิเคราะห์นี้ได้รับการแนะนำจากทางกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทรา โดยที่

x_1 คือ Total Protein (TP) ที่มีค่าปกติอยู่ในช่วง 6 – 8 g%
x_2 คือ Albumin (Alb) ที่มีค่าปกติอยู่ในช่วง 3.4 – 5.5 g%
x_3 คือ Globulin (Glb) ที่มีค่าปกติอยู่ในช่วง 2 – 3.5 g%
x_4 คือ Total Bilirubin (TB) ที่มีค่าปกติอยู่ในช่วง 0.1 – 1.1 mg%
x_5 คือ Direct Bilirubin (DB) ที่มีค่าปกติอยู่ในช่วง 0 – 0.5 mg%
x_6 คือ SGOT ที่มีค่าปกติอยู่ในช่วง 0 – 45 U/L
x_7 คือ SGPT ที่มีค่าปกติอยู่ในช่วง 9 – 72 U/L
x_8 คือ Alkaline Phosphatase (ALP) ที่มีค่าปกติอยู่ในช่วง 38 – 126 U/L

1.2 กำหนดตัวอย่างเพื่อใช้เป็นตัวแทนกลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และ กลุ่มผู้ป่วยที่ไม่มีแนวโน้มเป็นโรคตับ โดยกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทราจะเป็นผู้จำแนกผู้ป่วยทั้ง 2 กลุ่มนี้ให้ในเบื้องต้น

1.3 ตรวจสอบข้อมูลว่ามีความสอดคล้องกับข้อกำหนดเบื้องต้นของการวิเคราะห์จำแนกกลุ่มหรือไม่ ดังนี้

1.3.1 ตรวจสอบตัวแปรอิสระทั้ง 8 ตัว ว่ามีการแจกแจงแบบปกติพหุหรือไม่

Barbara and Linda (2001) กล่าวว่า เมื่อตัวแปรอิสระที่นำมาศึกษามีทั้งตัวแปรต่อเนื่อง และตัวแปรเชิงกลุ่ม จึงเป็นเรื่องยากที่ข้อมูลจะมีการแจกแจงแบบปกติหลายตัวแปร แต่ถ้าจำนวนตัวอย่างที่ใช้ในงานวิจัยมีขนาดใหญ่ และวัตถุประสงค์หลักของงานวิจัยคือ ต้องการสร้างสมการจำแนกประเภทที่มีประสิทธิภาพในการจำแนกกลุ่มข้อมูลให้ถูกต้องมากที่สุด ดังนั้นข้อสมมติดังกล่าวจึงไม่ใช่ข้อสมมติเบื้องต้นที่จำเป็น

1.3.2 ตรวจสอบว่าความแปรปรวนร่วมของตัวแปรอิสระในแต่ละกลุ่มเท่ากันหรือไม่ หรือทดสอบว่า $H_0 : \Sigma_1 = \Sigma_2$

ถ้าไม่เป็นไปตามเงื่อนไขที่กำหนดทั้ง 2 ข้อนี้ จะต้องทำการปรับเปลี่ยนวิธีในการสร้างสมการจำแนกกลุ่มจาก Linear Discriminant Function มาใช้ Quadratic Discriminant Function

1.4 สร้างตัวแบบจำแนกกลุ่มโดยใช้ข้อมูลผลการตรวจการทำงานของตับครั้งแรกของผู้ป่วยโรงพยาบาลเมือง ฉะเชิงเทรา โดยสร้างสมการเชิงเส้นที่เป็นสมการแสดงความสัมพันธ์ระหว่างตัวแปรแบ่งกลุ่ม (Y) กับตัวแปรอิสระ หรือประมาณค่า เพื่อตรวจสอบความสัมพันธ์ของตัวแปรอิสระ โดยการคำนวณค่า β ที่ทำให้ค่าอัตราส่วนระหว่างความผันแปรระหว่างกลุ่ม กับความผันแปรภายในกลุ่มมีค่ามากที่สุด หรือมีเปอร์เซ็นต์การจัดผิดกลุ่มน้อยที่สุด สำหรับฟังก์ชันการจำแนกกลุ่มจะอยู่ในรูปเชิงเส้น ดังนี้

$$\hat{Y}_i = a + b_0 + b_1 x_1 + ... + b_8 x_8$$

โดยที่ $\hat{Y}_i$ คือ ตัวแปรตาม หรือเรียกว่า Discriminant Score

b_i คือ สัมประสิทธิ์ของสมการจำแนกกลุ่ม

x_i คือ ตัวแปรอิสระที่ i เมื่อ i = 1, 2, …, 8 หรือเรียกว่าตัวแปรจำแนกกลุ่ม (Discriminator Variable)

ซึ่งจากงานวิจัยนี้จะได้ตัวแบบจำแนกกลุ่มจำนวน 2 สมการ เพื่อใช้สำหรับแบ่งกลุ่มผู้ป่วยออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็น และไม่เป็นโรคตับ

1.5 นำตัวแบบ หรือฟังก์ชันจำแนกกลุ่มที่ได้จากข้างต้นมาตรวจสอบกับข้อมูลชุดใหม่ โดยการแทนค่าตัวแปรอิสระทั้ง 8 ตัวในตัวแบบ

1.6 นำตัวแบบที่ได้รับการตรวจสอบแล้วมาใช้ในการพยากรณ์กลุ่มผู้ป่วยที่ได้รับการตรวจการทำงานของตับเป็นนครั้งแรกของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราในอนาคต

2. การวิเคราะห์การถดถอยแบบโลจิสติก (Logistic Regression Analysis) ประกอบด้วยขั้นตอนดังต่อไปนี้

2.1 กำหนดตัวแปร x_i ; i = 1, 2, …, 8 เป็นตัวแปรอิสระที่คาดว่าจะทำให้เกิดความแตกต่างระหว่างกลุ่มเช่นเดียวกับการวิเคราะห์จำแนกกลุ่ม

2.2 กำหนดตัวอย่างเพื่อใช้เป็นตัวแทนกลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และ กลุ่มผู้ป่วยที่ไม่มีแนวโน้มเป็นโรคตับ โดยกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทราจะเป็นผู้จำแนกผู้ป่วยทั้ง 2 กลุ่มนี้ให้ในเบื้องต้น

2.3 ตรวจสอบข้อมูลว่าสอดคล้องดับข้อกำหนดเบื้องต้นของการวิเคราะห์การถดถอย โลจิสติกหรือไม่ ดังนี้

2.3.1 ตรวจสอบชนิดของข้อมูลของตัวแปรอิสระ ซึ่งอาจเป็นข้อมูลที่มีได้ 2 ค่า หรือเป็นสเกลอันตรภาค (Interval Scale) หรือสเกลอัตราส่วน (Ratio Scale) ก็ได้

2.3.2 ค่าคาดหวังของค่าคาดเคลื่อนเป็นศูนย์ หรือ E(e) = 0

2.3.3 e_i และ e_j เป็นอิสระกัน; i, j = 1, 2, …, n (เมื่อ n คือ ขนาดตัวอย่างทั้งหมด)

2.3.4 e_i และ x_i เป็นอิสระกัน; i = 1, 2, …, n (เมื่อ n คือ ขนาดตัวอย่างทั้งหมด)

2.3.5 ตัวแปรอิสระเป็นอิสระต่อกัน

สำหรับกรณีที่ข้อมูลไม่เป็นไปตามข้อกำหนดเบื้องต้นให้ทำการแปลงข้อมูลก่อนที่จะนำไปสร้างสมการพยากรณ์

2.4 สร้างสมการสำหรับการพยากรณ์ ดังนี้

$$P(\text{เกิดเหตุการณ์}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

$$P(\text{ไม่เกิดเหตุการณ์}) = 1 - P(\text{เกิดเหตุการณ์})$$

แล้วตรวจสอบความถูกต้องเหมาะสมของสมการ โดยพิจารณาจากค่าสถิติ Hosmer และ Lemeshow และค่าสถิติวอลด์

2.5 นำตัวแบบที่ได้จากข้างต้นมาตรวจสอบกับข้อมูลชุดใหม่ โดยการแทนค่าตัวแปรอิสระทั้ง 8 ตัวในตัวแบบ

2.6 จากตัวแบบการถดถอยโลจิสติกนำไปพยากรณ์กลุ่มผู้ป่วยที่ได้รับการตรวจการทำงานของตับเป็นนครั้งแรกสำหรับผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราในอนาคต

3. เปรียบเทียบอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate : APER)

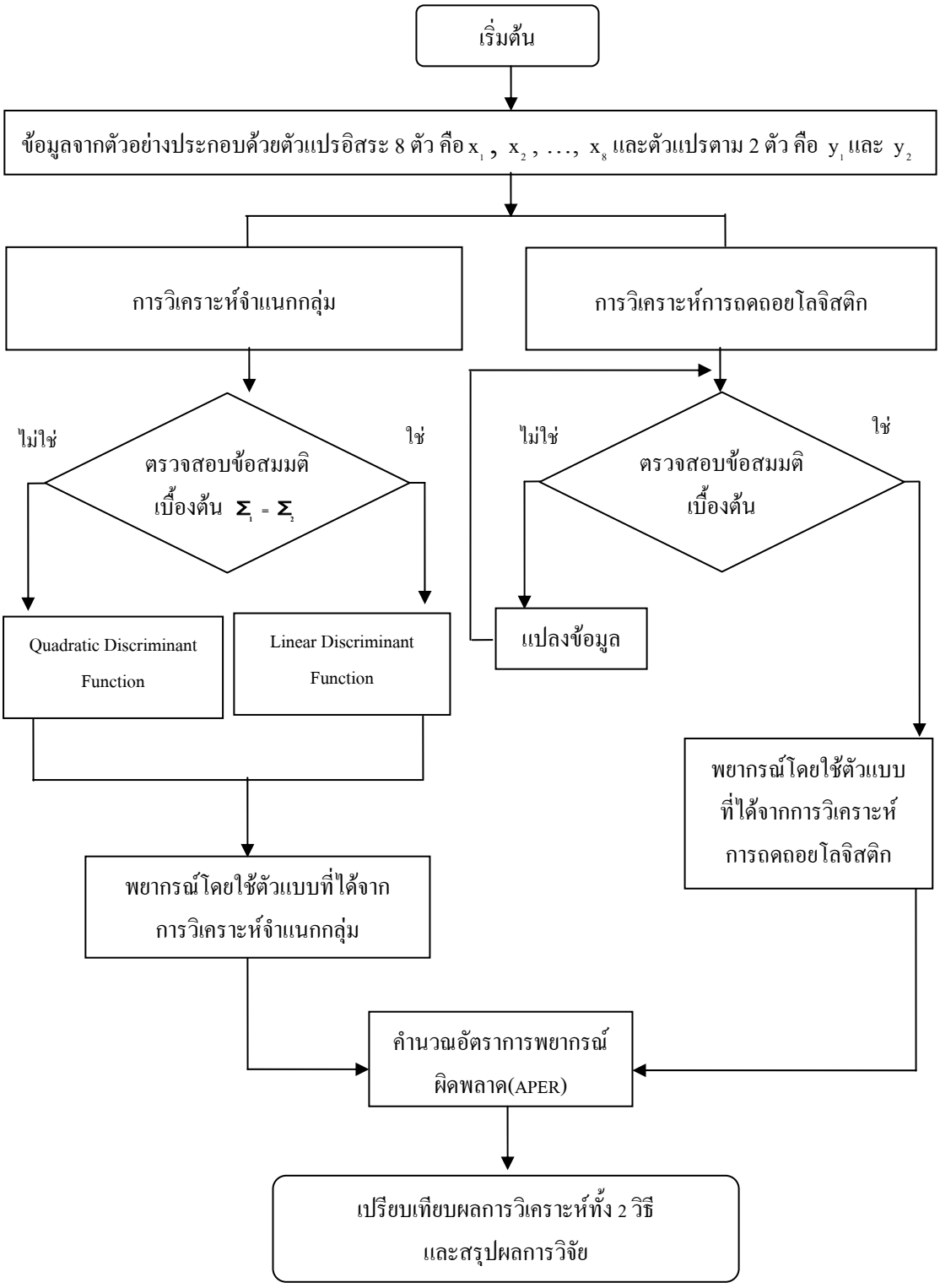
เปรียบเทียบผลการพยากรณ์ของทั้ง 2 วิธี คือ วิธีการวิเคราะห์จำแนกกลุ่ม และวิธีการวิเคราะห์การถดถอยแบบโลจิสติกด้วยอัตราความผิดพลาดในการพยากรณ์ได้จากอัตราความผิดพลาดหรือสัดส่วนของผู้ป่วยในกลุ่มตัวอย่างที่พยากรณ์ผิดพลาด คือ

$$\text{อัตราความผิดพลาด (APER)} = \frac{\text{จำนวนผู้ป่วยทั้งหมดที่พยากรณ์ผิดพลาด}}{\text{จำนวนผู้ป่วยทั้งหมดในกลุ่มตัวอย่าง}}$$

เมื่ออัตราความผิดพลาด (APER) มีค่าน้อย แสดงว่ามีประสิทธิภาพในการพยากรณ์สูง ในทางกลับกัน ถ้าอัตราความผิดพลาดมีค่ามาก แสดงว่ามีประสิทธิภาพในการพยากรณ์ต่ำ ซึ่งถ้าวิธีการใดให้ค่าอัตราความผิดพลาดน้อยกว่า แสดงว่าวิธีการนั้นมีประสิทธิภาพในการพยากรณ์มากกว่า

สถานที่ และระยะเวลาทำการวิจัย

สถานที่ทำการวิจัย ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน มีระยะเวลาการทำวิจัยตั้งแต่เดือน พฤศจิกายน 2548 ถึง กันยายน 2550



ภาพที่ 3 ขั้นตอนการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์ถดถอยโลจิสติก

ผลและวิจารณ์

ผล

งานวิจัยนี้ได้ทำการศึกษาการวินิจฉัยการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราจากกลุ่มผู้ป่วย 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มการเป็นโรคตับ และกลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับกลุ่มละ 150 ราย ซึ่งทางกลุ่มงานเทคนิคการแพทย์ โรงพยาบาลเมืองฉะเชิงเทราเป็นผู้จำแนกผู้ป่วยโรคตับนี้ โดยพิจารณาจากตัวแปรอิสระทั้ง 8 ตัวที่นำมาทำการศึกษาในงานวิจัยนี้

1. Total protein (TP: x_1): โปรตีนนี้สร้างในตับ โดยมีค่าปกติอยู่ในช่วง 6 – 8 g%
2. Albumin (Alb: x_2): เป็นโปรตีนที่สร้างจากตับ มีค่าปกติอยู่ในช่วง 3.4 – 5.5 g%
3. Globulin (Glb: x_3): เป็นโปรตีนกลุ่มหนึ่งนอกเหนือจากอัลบูมิน ค่านี้ได้จากการนำค่า Total Protein ลบด้วยค่า Albumin โดยมีค่าปกติอยู่ในช่วง 2 – 3.5 g%
4. Total Bilirubin (TB: x_4): เป็นน้ำดีชนิดหนึ่ง โดยมีค่าปกติอยู่ในช่วง 0.1 – 1.1 mg%
5. Direct Bilirubin(DB: x_5): เป็นน้ำดีชนิดหนึ่ง โดยมีค่าปกติอยู่ในช่วง 0 – 0.5 mg%
6. SGOT (x_6): เป็นเอนไซน์ในตับที่ช่วยในการสร้างกรดอะมิโน และโปรตีนโดยมีค่าปกติอยู่ในช่วง 0 – 45 U/L
7. SGPT (x_7): เป็นเอนไซน์ในตับที่มีค่าปกติอยู่ในช่วง 9 – 72 U/L
8. Alkaline Phosphatase (x_8): เป็นสารเอ็นไซม์ที่สร้างจากตับและกระดูกโดยมีค่าปกติอยู่ในช่วง 38 – 126 U/L

ซึ่งในงานวิจัยนี้จะทำการศึกษาวิเคราะห์ข้อมูลโดยใช้เทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก เพื่อให้ได้ตัวแบบที่ใช้ในการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย พร้อมทั้งเปรียบเทียบประสิทธิภาพของการวิเคราะห์ทั้ง 2 วิธี โดยการพิจารณาค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate : APER) โดยวิธีที่มีประสิทธิภาพสูงกว่าจะเป็นวิธีที่ให้ค่า APER ต่ำกว่า

ขอบเขตในการศึกษาครั้งนี้เฉพาะกลุ่มตัวอย่างผู้ป่วยในของโรงพยาบาลเมืองฉะเชิงเทรา ที่ได้รับการตรวจการทำงานของตับเป็นครั้งแรก ในช่วงระหว่างเดือนกันยายน 2548 – ตุลาคม 2549 จำนวน 300 ราย และมีการจำแนกกลุ่มผู้ป่วยนี้ออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มการเป็นโรคตับ และผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับกลุ่มละ 150 ราย ซึ่งการจำแนกกลุ่มในเบื้องต้นนี้ ได้รับการอนุเคราะห์จำแนกกลุ่มโดยกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทรา

ผลการวิจัยจะสรุปออกเป็น 4 ส่วนใหญ่ ๆ คือ

1. ผลการวิเคราะห์ข้อมูลเบื้องต้น
2. ผลการวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis)
3. ผลการวิเคราะห์การถดถอยแบบโลจิสติก (Logistic Regression Analysis)
4. ผลการเปรียบเทียบประสิทธิภาพของการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก

1. ผลการวิเคราะห์ข้อมูลเบื้องต้น

ตารางที่ 5 แสดงข้อมูลของตัวแปรอิสระทั้ง 8 ตัว ที่ทำการศึกษาจากขนาดตัวอย่าง 300 ราย

ตัวแปร	ค่าปกติ	ค่าเฉลี่ย (Mean)	ค่ามัธยฐาน (Median)	ค่าฐานนิยม (Mode)	ค่าสูงสุด (Max)	ค่าต่ำสุด (Min)	ส่วนเบี่ยงเบนมาตรฐาน (SD.)
x_1	6 - 8 g%	7.110	7.2	7.2	10.2	4.1	0.9607
x_2	3.4 - 5.5 g%	3.700	3.9	4.3	5.0	1.3	0.7909
x_3	2 - 3.5 g%	3.408	3.3	3.3	5.6	1.6	0.7178
x_4	0.1 - 1.1 mg%	1.013	0.7	0.5	8.4	0.1	1.1788
x_5	0 - 0.5 mg%	0.371	0.1	0.0	7.5	0.0	0.9260
x_6	0 - 45 U/L	62.161	41	24	321.0	0.3	54.1587
x_7	9 - 72 U/L	46.127	35	26	283.0	5.0	38.6569
x_8	38 – 126 U/L	130.477	101	83	762.0	43	93.4157

โดยค่าปกติของตัวแปรอิสระทั้ง 8 ตัว หมายถึง ค่า Cut Point ซึ่งเป็นค่าที่ได้จากการพิจาณา Total protein (TP), Albumin (Alb), Globulin (Glb), To tal Bilirubin (TB), Direct Bilirubin(DB), SGOT, SGPT และ Alkaline Phosphatase (ALK) ในคนปกติว่าโดยรวมแล้วมีค่าอยู่ในระดับเท่าไร สำหรับแต่ละตัวแปร

2. ผลการวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis)

ในงานวิจัยนี้ได้มีการแบ่งกลุ่มผู้ป่วยโดยพิจารณาจากตัวแปรอิสระทั้ง 8 ตัว ออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และมีแนวโน้มไม่เป็นโรคตับ โดยมีตัวแปรอิสระ (x_i ; i = 1, 2, …, 8) ที่แพทย์คาดว่าจะทำให้เกิดความแตกต่างระหว่างกลุ่มจำนวน 8 ตัวแปร ซึ่งผลจากการวิเคราะห์นี้ทำให้ได้ตัวแบบที่ใช้ในการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย

2.1 ผลการตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์จำแนกกลุ่ม

2.1.1 การตรวจสอบว่าเวกเตอร์ตัวแปรอิสระ **x** มีการแจกแจงแบบปกติพหุ(Multivariate Normal Distribution) หรือไม่

Barbara and Linda (2001) กล่าวว่า เมื่อตัวแปรอิสระที่นำมาศึกษามีทั้งตัวแปรต่อเนื่อง และตัวแปรเชิงกลุ่ม จึงเป็นเรื่องยากที่ข้อมูลจะมีการแจกแจงแบบปกติหลายตัวแปร แต่ถ้าจำนวนตัวอย่างที่ใช้ในงานวิจัยมีขนาดใหญ่ และวัตถุประสงค์หลักของงานวิจัยคือ ต้องการสร้างสมการจำแนกประเภทที่มีประสิทธิภาพในการจำแนกกลุ่มข้อมูลให้ถูกต้องมากที่สุด ดังนั้นข้อสมมติดังกล่าวจึงไม่ใช่สิ่งจำเป็น

2.1.2 การตรวจสอบการเท่ากันของเมทริกซ์ความแปรปรวนร่วมโดยวิธี Box'M ให้ผลการศึกษาดังนี้

สมมติฐานในการทดสอบ $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ และ $H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

หรือ H_0 : เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระทั้ง 2 กลุ่มเท่ากัน และ

H_1 : เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระทั้ง 2 กลุ่มไม่เท่ากัน

ตารางที่ 6 แสดงผลการทดสอบด้วยวิธี Box's M และประมาณค่าด้วยสถิติไคกำลังสอง

Chi – Square	DF	Pr > Chisq
1591.95	36	< 0.0001

ผลการวิเคราะห์โดยวิธี Box' M ที่ประมาณค่าด้วยไคกำลังสอง พบว่า ค่าสถิติ Chi – Square เท่ากับ 1591.95 และค่า P – Value น้อยกว่า 0.0001 ที่ระดับนัยสำคัญ 0.05 ดังนั้นจึงสรุปได้ว่าข้อมูลทั้ง 2 กลุ่มมีเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระไม่เท่ากันอย่างมีนัยสำคัญทางสถิติ

จากผลการทดสอบเกี่ยวกับเมทริกซ์ความแปรปรวนร่วม พบว่า ไม่สอดคล้องกับข้อสมมติฐานของการวิเคราะห์จำแนกกลุ่ม ดังนั้นลักษณะของข้อมูลตัวอย่างในการศึกษานี้จึงไม่เหมาะสมที่จะใช้ฟังก์ชันจำแนกกลุ่มเชิงเส้นตรงในการพยากรณ์ และอาจแก้ปัญหาโดยการใช้ฟังก์ชันจำแนกกลุ่มกำลังสองในการสร้างตัวแบบพยากรณ์กลุ่มผู้ป่วยดังต่อไปนี้

2.2 ผลการวิเคราะห์จำแนกกลุ่ม

จากข้อมูลที่ใช้ในการวิเคราะห์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา ที่ได้รับการตรวจการทำงานของตับ (Liver Function Test) จำนวน 300 ราย ที่มีตัวแปรอิสระที่ศึกษาทั้งสิ้น 8 ตัวแปร พบว่า ฟังก์ชันจำแนกกลุ่มกำลังสองมีดังต่อไปนี้

กลุ่มที่ 1 (กลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ)

$$\begin{aligned}\hat{Y}_1 = {} & -109.090 + 12.996x_1 + 11.447x_2 + 12.969x_3 + 16.128x_4 - 4.460x_5 \\ & + 0.238x_6 - 0.015x_7 + 0.245x_8 + 105.718x_1x_2 + 105.896x_1x_3 \\ & + 0.158x_1x_4 - 1.152x_1x_5 + 0.114x_1x_6 - 0.052x_1x_7 - 0.018x_1x_8 \\ & - 103.452x_2x_3 - 0.350x_2x_4 - 3.432x_2x_5 - 0.154x_2x_6 + 0.102x_2x_7 \\ & - 0.002x_2x_8 - 1.702x_3x_4 + 6.850x_3x_5 - 0.050x_3x_6 + 0.002x_3x_7 \\ & + 0.018x_3x_8 + 13.890x_4x_5 + 0.048x_4x_6 + 0.018x_4x_7 - 0.040x_4x_8 \\ & - 0.158\, x_5x_6 + 0.006x_5x_7 + 0.126x_5x_8 + 0.006x_6x_7 - 53.741x_1^2 \\ & - 54.977x_2^2 - 56.769x_3^2 - 8.797x_4^2 - 51.671x_5^2 - 0.008x_6^2 - 0.005x_7^2 \\ & - 0.001x_8^2\end{aligned}$$

กลุ่มที่ 2 (กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ)

$$\begin{aligned}\hat{Y}_2 = {} & -27.429 - 1.910x_1 + 6.803x_2 + 6.263x_3 + 1.581x_4 - 0.866x_5 \\ & + 0.018x_6 - 0.006x_7 + 0.006x_8 + 112.122x_1x_2 + 112.296x_1x_3 \\ & + 0.608x_1x_4 - 0.778x_1x_5 + 0.012x_1x_6 - 0.002x_1x_7 + 0.002x_1x_8 \\ & - 112.030x_2x_3 - 0.306x_2x_4 + 0.060x_2x_5 - 0.014x_2x_6 + 0.006x_2x_7 \\ & - 0.002x_2x_8 - 0.484x_3x_4 + 0.612x_3x_5 - 0.006x_3x_6 - 0.002x_3x_7 \\ & + 0.002x_3x_8 + 3.392x_4x_5 - 0.006x_4x_6 + 0.002x_4x_7 - 0.004x_4x_8 \\ & + 0.014x_5x_6 - 0.004x_5x_7 + 0.006x_5x_8 - 56.158x_1^2 - 56.767x_2^2 \\ & - 56.947x_3^2 - 1.413x_4^2 - 2.460x_5^2\end{aligned}$$

โดยที่

x_1 คือ Total Protein (TP)

x_2 คือ Albumin (Alb)

x_3 คือ Globulin (Glb)

x_4 คือ Total Bilirubin (TB)

x_5 คือ Direct Bilirubin (DB)

x_6 คือ SGOT

x_7 คือ SGPT

x_8 คือ Alkaline Phosphatase (ALP)

สำหรับการพยากรณ์โดยฟังก์ชันจำแนกกลุ่มกำลังสองของหน่วยตัวอย่างใหม่นั้นทำได้โดยแทนค่าตัวแปรทั้ง 8 ลงในสมการ $\hat{Y}_1$ และ $\hat{Y}_2$ โดยมีเกณฑ์ในการพิจารณาเพื่อจัดกลุ่มผู้ป่วยดังนี้

ถ้า $\hat{Y}_1 > \hat{Y}_2$ จะพยากรณ์ให้ผู้ป่วยอยู่ในกลุ่มที่ 1 (ผู้ป่วยมีแนวโน้มไม่เป็นโรคตับ)

ถ้า $\hat{Y}_1 < \hat{Y}_2$ จะพยากรณ์ให้ผู้ป่วยอยู่ในกลุ่มที่ 2 (ผู้ป่วยมีแนวโน้มเป็นโรคตับ)

2.3 ผลการทดสอบสมมติฐานของการวิเคราะห์จำแนกกลุ่ม

ในการวิเคราะห์จำแนกกลุ่มนั้นผู้วิจัยต้องทราบจำนวนกลุ่มของข้อมูลในเบื้องต้นก่อน ดังนั้นทางกลุ่มงานเทคนิคการแพทย์ โรงพยาบาลเมืองฉะเชิงเทรา จึงได้แบ่งกลุ่มของผู้ป่วยขั้นต้น โดยการแบ่งผู้ป่วยจำนวนทั้งสิ้น 300 ราย ออกเป็น 2 กลุ่มคือ กลุ่มที่ 1 มีแนวโน้มที่จะเป็นโรคตับ จำนวน 150 ราย และกลุ่มที่ 2 มีแนวโน้มที่จะไม่เป็นโรคตับจำนวน 150 ราย ซึ่งทางกลุ่มงานเทคนิคการแพทย์ โรงพยาบาลเมืองฉะเชิงเทราอาศัยการพิจารณาจากค่าปกติของทั้ง 8 ตัวแปร ดังนั้นขนาดตัวอย่างของกลุ่มที่ 1 : กลุ่มที่ 2 เท่ากับ 0.5 : 0.5 ซึ่งการแบ่งกลุ่มนี้ควรทำการตรวจสอบก่อนว่าทั้ง 2 กลุ่มนี้แตกต่างกันจริงหรือไม่ ก่อนที่จะทำการวิเคราะห์ในขั้นต่อไป ซึ่งในงานวิจัยนี้ได้เลือกใช้วิธี Hotelling T^2 ในการทดสอบดังนี้

สมมติฐานในการทดสอบ $H_0 : \mu_1 = \mu_2$ และ $H_1 : \mu_1 \neq \mu_2$
หรือ H_0 : ค่าเฉลี่ยของกลุ่มที่ 1 เท่ากับกลุ่มที่ 2 และ
H_1 : ค่าเฉลี่ยของกลุ่มที่ 1 ไม่เท่ากับกลุ่มที่ 2

ตารางที่ 7 แสดงผลการทดสอบสถิติ Hotelling T^2 และประมาณค่าด้วยสถิติ F

Hotelling T^2	F Value	Num DF	Den DF	Pr > F
1.4323	52.10	8	291	< 0.0001

ผลการวิเคราะห์โดยวิธี Hotelling T^2 พบว่า สถิติ Hotelling T^2 เท่ากับ 1.4323 และค่า P – Value น้อยกว่า 0.0001 ที่ระดับนัยสำคัญ 0.05 จึงสรุปได้ว่าข้อมูลทั้ง 2 กลุ่มมีค่าเฉลี่ยที่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ หรือตัวแบบที่สร้างสามารถนำไปใช้ในการจำแนกกลุ่ม

2.4 ผลการประเมินความเหมาะสมของฟังก์ชันจำแนกกลุ่ม

$$\hat{Y}_1 = -109.090 + 12.996x_1 + 11.447x_2 + 12.969x_3 + 16.128x_4 - 4.460x_5 + 0.238x_6 - 0.015x_7 + 0.245x_8 + 105.718x_1x_2 + 105.896x_1x_3 + 0.158x_1x_4 - 1.152x_1x_5 + 0.114x_1x_6 - 0.052x_1x_7 - 0.018x_1x_8 - 103.452x_2x_3 - 0.350x_2x_4 - 3.432x_2x_5 - 0.154x_2x_6 + 0.102x_2x_7 - 0.002x_2x_8 - 1.702x_3x_4 + 6.850x_3x_5 - 0.050x_3x_6 + 0.002x_3x_7 + 0.018x_3x_8 + 13.890x_4x_5 + 0.048x_4x_6 + 0.018x_4x_7 - 0.040x_4x_8 - 0.158\ x_5x_6 + 0.006x_5x_7 + 0.126x_5x_8 + 0.006x_6x_7 - 53.741x_1^2 - 54.977x_2^2 - 56.769x_3^2 - 8.797x_4^2 - 51.671x_5^2 - 0.008x_6^2 - 0.005x_7^2 - 0.001x_8^2$$

$$\hat{Y}_2 = -27.429 - 1.910x_1 + 6.803x_2 + 6.263x_3 + 1.581x_4 - 0.866x_5 + 0.018x_6 - 0.006x_7 + 0.006x_8 + 112.122x_1x_2 + 112.296x_1x_3 + 0.608x_1x_4 - 0.778x_1x_5 + 0.012x_1x_6 - 0.002x_1x_7 + 0.002x_1x_8 - 112.030x_2x_3 - 0.306x_2x_4 + 0.060x_2x_5 - 0.014x_2x_6 + 0.006x_2x_7 - 0.002x_2x_8 - 0.484x_3x_4 + 0.612x_3x_5 - 0.006x_3x_6 - 0.002x_3x_7 + 0.002x_3x_8 + 3.392x_4x_5 - 0.006x_4x_6 + 0.002x_4x_7 - 0.004x_4x_8 + 0.014x_5x_6 - 0.004x_5x_7 + 0.006x_5x_8 - 56.158x_1^2 - 56.767x_2^2 - 56.947x_3^2 - 1.413x_4^2 - 2.460x_5^2$$

การประเมิน หรือตัดสินใจว่าจะนำฟังก์ชันจำแนกกลุ่มที่ได้ไปใช้หรือไม่จะพิจารณาจากค่าร้อยละของการพยากรณ์ หรือการจัดกลุ่มได้ถูกต้อง ซึ่งในงานวิจัยนี้ได้ใช้วิธีการ Cross – Validation เพื่อประเมินความเหมาะสมของฟังก์ชันจำแนกกลุ่ม โดยมีหลักเกณฑ์ในการประเมินคือ จะใช้ข้อมูล 299 ค่า มาสร้างสมการจำแนกประเภท แล้วนำสมการที่สร้างขึ้นไปพยากรณ์ หรือจัดกลุ่มหน่วยที่เหลืออีก 1 หน่วย ให้ทำเช่นนี้ไป 300 ครั้ง แล้วนับจำนวนการพยากรณ์กลุ่มที่ผิด ซึ่งมีรายละเอียดดังนี้

ครั้งที่ 1 ใช้ข้อมูลจำนวน 299 หน่วย คือ หน่วยที่ 2 ถึงหน่วยที่ 300 มาสร้างสมการจำแนกประเภท แล้วนำสมการที่สร้างขึ้นมาพยากรณ์กลุ่มให้กับข้อมูลหน่วยที่ 1 แล้วทำการตรวจสอบว่าพยากรณ์ถูกหรือผิด

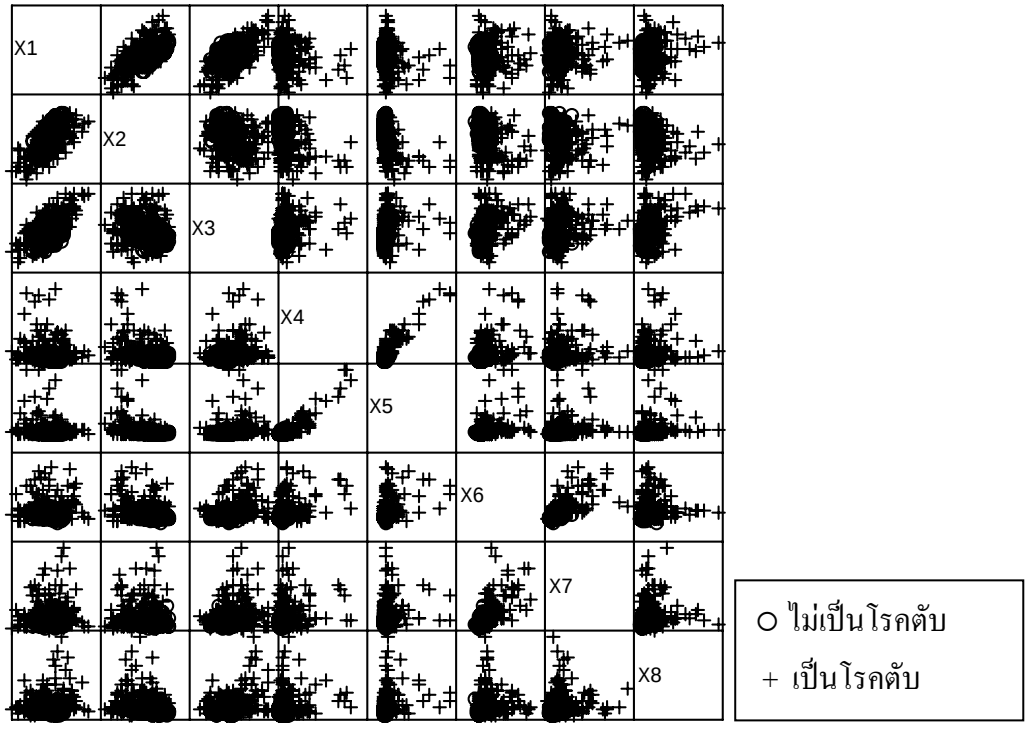
ครั้งที่ 2 ใช้ข้อมูลจำนวน 299 หน่วย คือ หน่วยที่ 1 และหน่วยที่ 3 ถึงหน่วยที่ 300 มาสร้างสมการจำแนกประเภท โดยเก็บหน่วยที่ 2 ไว้สำหรับตรวจสอบ แล้วนำผลวิเคราะห์มาพยากรณ์กลุ่มให้กับข้อมูลหน่วยที่ 2 แล้วทำการตรวจสอบว่าพยากรณ์ถูกหรือผิด

.

.

.

ครั้งที่ 300 ใช้ข้อมูลหน่วยที่ 1 ถึงหน่วยที่ 299 โดยเก็บหน่วยที่ 300 ไว้สำหรับตรวจสอบ มาสร้างตัวแบบจำแนกประเภทแล้วพยากรณ์กลุ่มให้กับข้อมูลหน่วยที่ 300 และตรวจสอบความถูกต้องของการพยากรณ์



ภาพที่ 4 แสดงกราฟระหว่างตัวแปรอิสระทั้ง 8 ตัว คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) ต่อแนวโน้มการเป็น หรือไม่เป็นโรคตับ

เมื่อนำข้อมูลผู้ป่วยทั้ง 300 ราย มาสร้างกราฟจะได้ดังรูปที่ 4 ซึ่งเห็นได้ว่ามีค่าสังเกตบางรายที่อยู่ในกลุ่มที่มีแนวโน้มไม่เป็นโรคตับ แต่ถูกจัดให้อยู่ในกลุ่มที่มีแนวโน้มเป็นโรคตับ และมีค่าสังเกตที่อยู่ในกลุ่มที่มีแนวโน้มเป็นโรคตับบางรายถูกจัดให้อยู่ในกลุ่มที่มีแนวโน้มไม่เป็นโรคตับ แสดงว่าในการจำแนกกลุ่มอาจมีความผิดพลาดเกิดขึ้น

ตารางที่ 8 แสดงผลการจำแนกกลุ่มโดยวิธีการจำแนกกลุ่ม

กลุ่ม	ขนาดตัวอย่าง (n_i)	ผลการจำแนกกลุ่ม	
		มีแนวโน้มไม่เป็นโรคตับ	มีแนวโน้มเป็นโรคตับ
มีแนวโน้มไม่เป็นโรคตับ	150	145 (96.67%)	5 (3.33%)
มีแนวโน้มเป็นโรคตับ	150	2 (1.33%)	148 (98.67%)

ดังนั้นอัตราของการจำแนกกลุ่มผิดพลาด = $\frac{(7\times100)}{300} = 2.33\%$

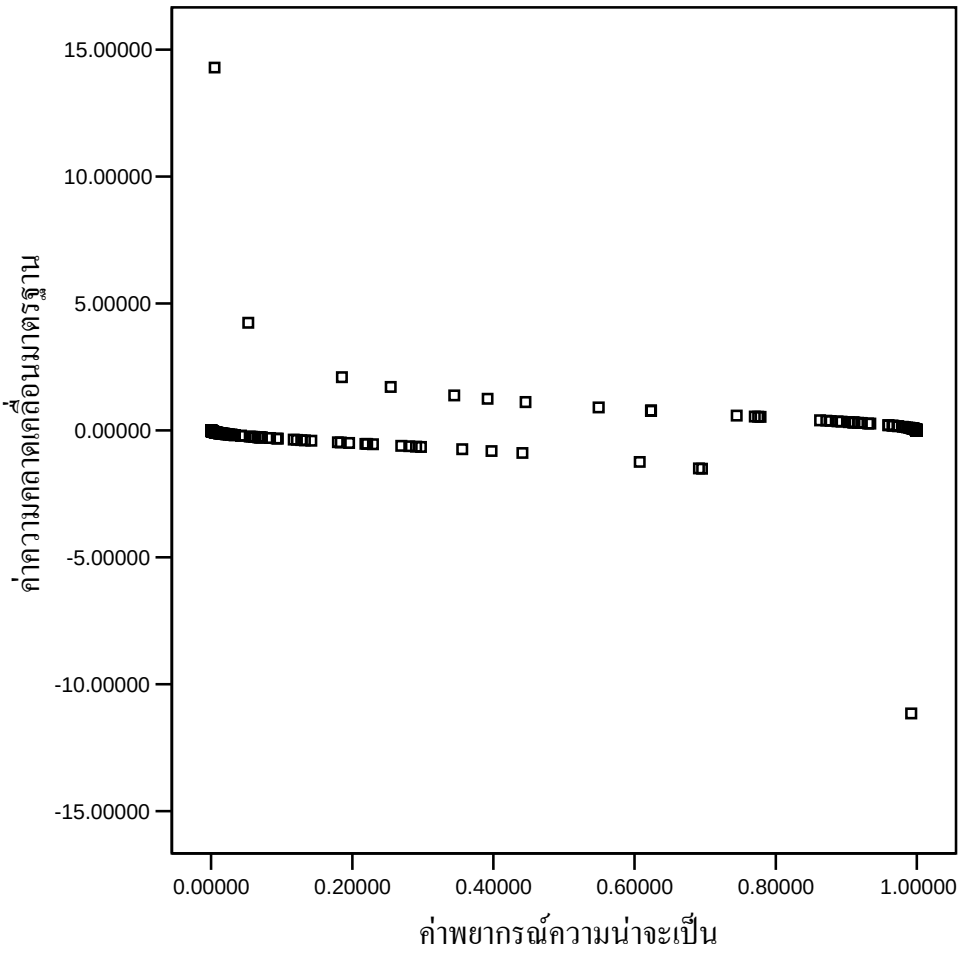
ผลการประเมินความเหมาะสมของฟังก์ชันจำแนกกลุ่มกำลังสอง พบว่า ความคลาดเคลื่อนของการพยากรณ์ หรือจำแนกกลุ่มมีค่าประมาณ 2.33% หรืออีกนัยหนึ่งกล่าวได้ว่าฟังก์ชันจำแนกกลุ่มกำลังสองนี้สามารถใช้ในการพยากรณ์ หรือจำแนกกลุ่มผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 97.67%

3. ผลการวิเคราะห์การถดถอยแบบโลจิสติก (Logistic Regression Analysis)

ในงานวิจัยนี้เป็นการพยากรณ์กลุ่มผู้ป่วยโดยออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และมีแนวโน้มไม่เป็นโรคตับ โดยมีตัวแปรอิสระที่คาดว่าจะทำให้เกิดความแตกต่างระหว่างกลุ่มจำนวน 8 ตัวแปร ซึ่งจากการวิเคราะห์นี้ทำให้ได้สมการที่ใช้ในการพยากรณ์แนวโน้มการเป็นหรือไม่เป็นโรคตับของผู้ป่วย โดยมีผลการวิเคราะห์ดังนี้

3.1 ผลการตรวจสอบเงื่อนไขของการวิเคราะห์การถดถอยแบบโลจิสติก

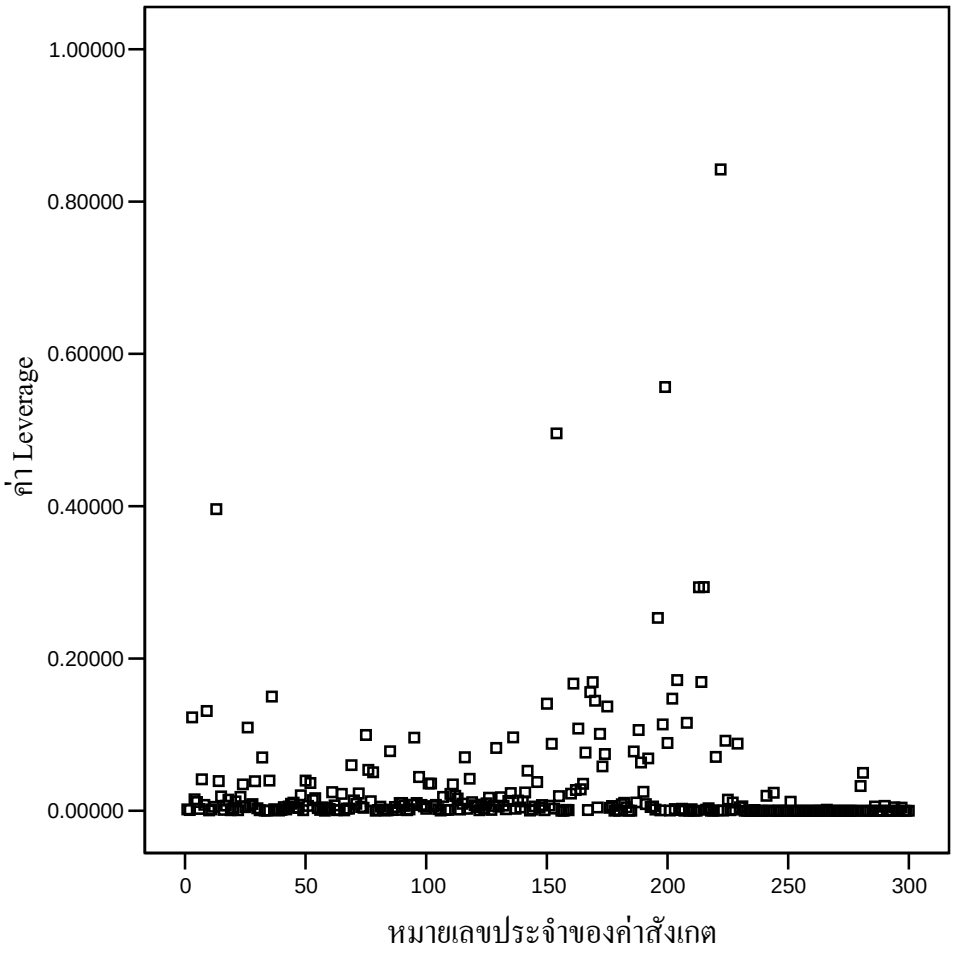
3.1.1 การตรวจสอบว่าค่าคาดหวังของค่าคลาดเคลื่อนเป็นศูนย์ หรือE(e) = 0



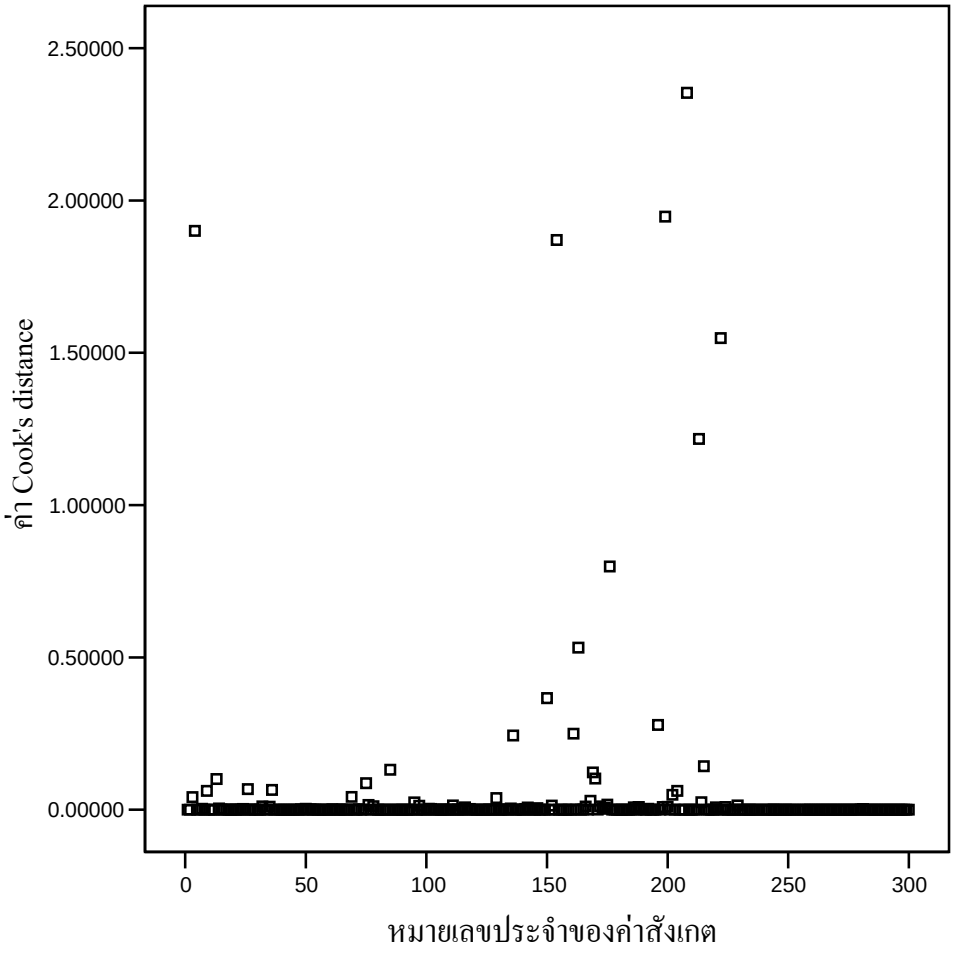
ภาพที่ 5 แสดงการกระจายของส่วนเหลือ (Standard Residual) กับค่าพยากรณ์ความน่าจะเป็น

จากภาพที่ 4 แสดงการพลอตการกระจายของส่วนเหลือ เพื่อตรวจสอบว่าตัวแบบที่นำมาใช้กับข้อมูลตัวอย่างสอดคล้องกับข้อตกลงเบื้องต้นที่ว่า E(e) = 0 ซึ่งจากการพิจารณารูปแบบการกระจายของส่วนเหลือ พบว่า ค่าส่วนเหลือมาตรฐานของค่าสังเกตส่วนใหญ่จะมีค่าอยู่รอบ ๆ ค่าศูนย์

3.1.2 การตรวจสอบว่า e_i และ e_j เป็นอิสระกัน หรือไม่ และตรวจสอบว่า e_i และ x_i เป็นอิสระกันหรือไม่; i, j = 1, 2, …, n (เมื่อ n คือ ขนาดตัวอย่างทั้งหมด) แสดงดังภาพที่ 5 และ 6



ภาพที่ 6 แสดงการกระจายของค่า Leverage กับหมายเลขประจำของค่าสังเกต



ภาพที่ 7 แสดงการกระจายของค่า Cook's distance กับหมายเลขประจำของค่าสังเกต

จากภาพที่ 6 และ 7 แสดงการพลอตระหว่างค่า Leverage และค่า Cook's กับหมายเลขประจำของค่าสังเกต ซึ่งผลการพิจารณาจากรูปแบบการกระจายของข้อมูล พบว่า ค่า Leverage และค่า Cook's ส่วนใหญ่จะมีค่าอยู่รอบ ๆ ค่าศูนย์ แสดงว่าข้อมูลชุดนี้ e_i และ e_j เป็นอิสระกัน เช่นเดียวกับ e_i และ x_i ที่เป็นอิสระกัน

3.1.3 การตรวจสอบว่าตัวแปรอิสระมีความสัมพันธ์กันหรือไม่โดยพิจารณาจากค่า Variance Inflation Factor (VIF) โดยมีหลักเกณฑ์ในการพิจารณาคือ ถ้าค่า VIF มากกว่า 10 แสดงว่าตัวแปรอิสระนั้น ๆ มีความสัมพันธ์กัน แสดงดังตารางที่ 9

ตารางที่ 9 แสดงค่า Variance Inflation Factor ของตัวแปรอิสระทั้ง 8 ตัวแปร

ตัวแปร	df	VIF
Intercept	1	0
x_1	**1**	**99.92214**
x_2	**1**	**69.17887**
x_3	**1**	**56.79271**
x_4	1	5.95257
x_5	1	6.40402
x_6	1	2.63630
x_7	1	2.04825
x_8	1	1.36282

เมื่อพิจารณาจากค่า VIF พบว่า มีตัวแปรอิสระ 3 ตัวมีค่า VIF มากกว่า 10 คือ x_1, x_2 และ x_3 ซึ่งมีค่า VIF เท่ากับ 99.922, 69.179 และ 56.793 ตามลำดับ ดังนั้นจึงสรุปได้ว่าตัวแปรอิสระทั้ง 3 ตัวนี้มีความสัมพันธ์กับตัวแปรอิสระอื่น ๆ หรือกล่าวอีกนัยหนึ่งได้ว่าเกิดปัญหา Multicollinearity

เนื่องจากเมื่อข้อมูลตัวแปรอิสระที่นำมาศึกษามีความสัมพันธ์กัน ซึ่งจะทำให้เกิดปัญหา Multicollinearity และอาศัยข้อมูลทางการแพทย์ที่ว่าตัวแปรอิสระทั้ง 3 ตัวมีความสัมพันธ์กันเนื่องจาก Total protein (TP: x_1), Albumin (Alb: x_2) และ Globulin (Glb: x_3) เป็นโปรตีนที่ตับสังเคราะห์เช่นเดียวกัน โดยในการพิจารณาทางการแพทย์นั้น หากค่าใดค่าหนึ่งมีค่าที่สูง ค่าของตัวแปรที่เหลืออีก 2 ตัวก็จะมีค่าสูงเช่นกัน หรือกล่าวอีกนัยหนึ่งว่าหากค่าใดค่าหนึ่งมีค่าที่ต่ำ ค่าของตัวแปรที่เหลืออีก 2 ตัวก็จะมีค่าต่ำเช่นกัน จึงต้องมีการรวมตัวแปรอิสระโดยมีการปรับขนาดของตัวแปรอิสระ ซึ่งในงานวิจัยนี้ได้ทำให้ค่าของตัวแปรเป็นค่ามาตรฐาน (Standardized Score) แล้วจึงนำข้อมูลของทั้ง 3 ตัวแปรมารวมกันเพื่อสร้างเป็นตัวแปรใหม่ ซึ่งเป็นวิธีการที่ใช้ในการแก้ไขปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กัน และจากงานวิจัยนี้ได้นำค่าของตัวแปร x_1, x_2 และ x_3 แปลงเป็นค่ามาตรฐานแล้วนำมารวมกันเป็นตัวแปรใหม่ขึ้นมาคือ z_{123} หลังจากนั้นนำตัวแปร z_{123}, x_4, x_5, x_6, x_7 และ x_8 มาตรวจสอบความสัมพันธ์อีกครั้ง และเมื่อพิจารณาจากค่า VIF อีกครั้งปรากฏว่าตัวแปรทุกตัวให้ค่า VIF น้อยกว่า 10 จึงสรุปได้ตัวแปรอิสระไม่มีความสัมพันธ์กัน หรือกล่าวอีกนัยหนึ่งได้ว่าไม่เกิดปัญหา Multicollinearity ดังนั้นจึงลดตัวแปรให้เหลือเพียง 6 ตัวแปร คือ z_{123}, x_4, x_5, x_6,

x_7 และ x_8 แล้วจึงนำข้อมูลที่ได้มีการแก้ไขปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันแล้วมาทำการวิเคราะห์การถดถอยแบบโลจิสติกต่อไป

ตารางที่ 10 แสดงค่า Variance Inflation Factor ภายหลังการปรับขนาดของตัวแปรอิสระ

ตัวแปร	df	VIF
Intercept	1	0
z_{123}	1	1.06613
x_4	1	5.93899
x_5	1	6.39674
x_6	1	2.28865
x_7	1	1.98597
x_8	1	1.10830

โดยที่ z_{123} หมายถึง ตัวแปร x_1, x_2 และ x_3 ที่ได้แปลงเป็นค่ามาตรฐานแล้วนำมารวมกันเป็นตัวแปรใหม่ (z_{123})

3.2 ผลการวิเคราะห์การถดถอยแบบโลจิสติก

จากการวิเคราะห์ข้อมูลผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทรา ที่ได้รับการตรวจการทำงานของตับจำนวน 300 ราย ที่มีตัวแปรอิสระที่ศึกษาทั้งสิ้น 6 ตัวแปร โดยวิธีการวิเคราะห์การถดถอยแบบโลจิสติก และทำการประมาณค่าพารามิเตอร์โดยวิธีภาวะน่าจะเป็นสูงสุด ได้ฟังก์ชันโลจิทดังนี้

ตารางที่ 11 แสดงค่าสัมประสิทธิ์ของตัวแปรอิสระ

ตัวแปร	df	b_i
Intercept	1	13.2738
z_{123}	1	0.1503
x_4	1	-3.7401
x_5	1	0.9760
x_6	1	-0.1154

ตารางที่ 11 (ต่อ)

ตัวแปร	df	b_i
x_7	1	0.00787
x_8	1	-0.0466

$$\hat{g}(x) = 13.2738 + 0.1503z_{123} - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8$$

หรือ

$$\hat{g}(x) = 13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8$$

โดยตัวแปรที่มีผลต่อการพยากรณ์แนวโน้มผู้ป่วยทั้ง 8 ตัวแปรคือ

x_1 คือ Total Protein (TP)

x_2 คือ Albumin (Alb)

x_3 คือ Globulin (Glb)

x_4 คือ Total Bilirubin (TB)

x_5 คือ Direct Bilirubin (DB)

x_6 คือ SGOT

x_7 คือ SGPT

x_8 คือ Alkaline Phosphatase (ALP)

สำหรับการพยากรณ์โดยใช้ตัวแบบ หรือสมการถดถอยที่ได้จากการวิเคราะห์ถดถอยโลจิสติก ทำได้โดยการแทนค่าตัวแปรอิสระทั้ง 8 ตัวลงในสมการแล้วนำค่า $\hat{g}(x)$ หรือ Logit ($\hat{P}$ (แนวโน้มไม่เป็นโรคตับ)) ที่ได้แทนลงในสมการ

$$P(\text{เกิดเหตุการณ์}) = \frac{e^{13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8}}{1 + e^{13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8}}$$

ค่าที่คำนวณได้นี้เป็นค่าที่สามารถพยากรณ์แนวโน้มการไม่เป็นโรคตับของผู้ป่วยได้

3.2.1 การทดสอบสมมติฐานเกี่ยวกับค่าสัมประสิทธิ์ความถดถอยโลจิสติคของตัวแปรอิสระแต่ละตัวโดยวิธีการทดสอบวอลด์ดังนี้

สมมติฐานในการทดสอบ $H_0 : \hat{\beta}_i = 0$; i = 1,2, … , p และ $H_1 : \hat{\beta}_i \neq 0$
หรือ H_0 : ตัวแปรอิสระแต่ละตัวในตัวแบบไม่มีความสัมพันธ์กับตัวแปรตาม
และ H_1 : ตัวแปรอิสระแต่ละตัวในตัวแบบมีความสัมพันธ์กับตัวแปรตาม

ตารางที่ 12 แสดงค่าสัมประสิทธิ์ความถดถอยโลจิสติกของตัวแปรอิสระ

ตัวแปร	b_i	$SE(b_i)$
Intercept	13.2738	1.9486
z_{123}	0.1503	0.1279
x_4	-3.7401	0.9523
x_5	0.9760	1.6726
x_6	-0.1154	0.0194
x_7	0.00787	0.0173
x_8	-0.0466	0.00851

ตารางที่ 13 แสดงการตรวจสอบค่าสัมประสิทธิ์ความถดถอยโลจิสติคของตัวแปรอิสระ

Chi-Square	df	Pr > ChiSq
53.9230	6	< 0.0001

ผลการวิเคราะห์ด้วยสถิติวอลด์ พบว่า ตัวสถิติวอลด์มีค่าเท่ากับ 53.9230 และค่า P – Value < 0.0001 ที่ระดับนัยสำคัญ 0.05 จึงสรุปได้ว่าค่าสัมประสิทธิ์ของแต่ละตัวแปรไม่เท่ากับศูนย์อย่างมีนัยสำคัญทางสถิติ แสดงว่าตัวแปรอิสระแต่ละตัวในตัวแบบมีความสัมพันธ์กับตัวแปรตาม

ส่วนการตรวจสอบความเหมาะสมของสมการความถดถอยโลจิสติกในงานวิจัยนี้ได้นำวิธีการทดสอบ Hosmer และ Lemeshow (H – L) ที่มีการแบ่งข้อมูลออกเป็น p กลุ่มย่อยให้ผลการศึกษาดังนี้

สมมติฐานในการทดสอบ H_0 : ตัวแบบเหมาะสมกับข้อมูล และ H_1 : ตัวแบบไม่เหมาะสมกับข้อมูล

ตารางที่ 14 แสดงการตรวจสอบความเหมาะสมของสมการความถดถอยโลจิสติค

Chi-Square	df	Pr > ChiSq
26.2799	8	0.0009

ผลการวิเคราะห์โดยวิธี Hosmer และ Lemeshow พบว่า สถิติ Chi – Square มีค่าเท่ากับ 26.2799 และค่า P – Value เท่ากับ 0.0009 ที่ระดับนัยสำคัญ 0.05 จึงสรุปได้ว่าตัวแบบไม่เหมาะสมกับข้อมูลอย่างมีนัยสำคัญทางสถิติ

เนื่องจากการวิธีวิเคราะห์การถดถอยโลจิสติกมีตัวแบบที่ไม่เหมาะสมที่จะใช้ในการพยากรณ์หน่วยตัวอย่างใหม่ ดังนั้นหากต้องการพยากรณ์หน่วยตัวอย่างใหม่สำหรับข้อมูลการตรวจสอบการทำงานของตับในครั้งแรกนั้น ควรใช้วิธีการวิเคราะห์จำแนกกลุ่มมากกว่าวิธีวิเคราะห์การถดถอยโลจิสติก

3.3 ผลการประเมินความเหมาะสมของสมการที่ใช้ในการพยากรณ์

การประเมินความเหมาะสมของสมการที่ใช้ในการพยากรณ์ โดยพิจารณาจากค่า Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) สามารถสรุปได้ดังนี้

ตารางที่ 15 แสดงผลการประเมินความเหมาะสมของสมการที่ใช้ในการพยากรณ์

กลุ่ม	ขนาดตัวอย่าง (n_i)	ผลการพยากรณ์	
		มีแนวโน้มไม่เป็นโรคตับ	มีแนวโน้มเป็นโรคตับ
มีแนวโน้มไม่เป็นโรคตับ	150	143 (95.33%)	7 (4.67%)
มีแนวโน้มเป็นโรคตับ	150	10 (6.67%)	140 (93.33%)

ดังนั้น อัตราการพยากรณ์ผิดพลาด = $\frac{(17\times100)}{300}$ = 5.67%

จากข้อมูลจริงของผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับจำนวน 150 ราย เมื่อทำการพยากรณ์โดยวิธีการถดถอยโลจิสติก แล้วพบว่า มีการพยากรณ์ผู้ป่วยผิดพลาดจำนวน 7 ราย นั่นคือพยากรณ์ผิด 4.67% ในขณะที่เมื่อใช้ข้อมูลจริงของผู้ป่วยที่มีแนวโน้มเป็นโรคตับจำนวน 150 ราย พบว่า มีการพยากรณ์ผิดจำนวน 10 ราย คิดเป็น 6.67% ดังนั้นจึงสามารถสรุปผลอัตราความผิดพลาดจากการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโดยรวมได้เท่ากับ 5.67%หรือกล่าวอีกนัยหนึ่งว่าตัวแบบนี้สามารถใช้ในการพยากรณ์ กลุ่มผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 94.33%

4. ผลการเปรียบเทียบประสิทธิภาพของการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก

ในการศึกษาแนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราที่ได้รับการตรวจการทำงานของตับโดยจะแบ่งเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และมีแนวโน้มไม่เป็นโรคตับ ซึ่งตัวแปรอิสระที่คาดว่าจะทำให้เกิดความแตกต่างระหว่างกลุ่มในงานวิจัยนี้มีจำนวน 8 ตัวแปร ซึ่งในงานวิจัยนี้ได้ทำการเปรียบเทียบผลการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยแบบโลจิสติก ซึ่งทำการประเมินประสิทธิภาพในการพยากรณ์ได้โดยการพิจารณาจากอัตราความผิดพลาดในการพยากรณ์ที่เห็นชัด หรือสัดส่วนของผู้ป่วยในกลุ่มตัวอย่างที่พยากรณ์ผิดพลาด ถ้าอัตราความผิดพลาด (APER) มีค่าน้อย แสดงว่ามีประสิทธิภาพในการพยากรณ์

สูง และในทางตรงกันข้ามถ้าอัตราความผิดพลาดมีค่ามาก แสดงว่าตัวแบบที่ได้จากการวิเคราะห์นั้นมีประสิทธิภาพในการพยากรณ์ต่ำ

ตารางที่ 16 แสดงผลการเปรียบเทียบอัตราความผิดพลาดของการวิเคราะห์จำแนกกลุ่ม (DA) และการวิเคราะห์การถดถอยแบบโลจิสติก (LA)

กลุ่ม	กลุ่มตัวอย่าง	ผลการจำแนกกลุ่ม			
		มีแนวไม่โน้มเป็นโรคตับ		มีแนวโน้มเป็นโรคตับ	
		DA	LA	DA	LA
มีแนวโน้มไม่เป็นโรคตับ	150	145	143	5	7
		(96.67%)	(95.33%)	(3.33%)	(4.67%)
มีแนวโน้มเป็นโรคตับ	150	2	10	148	140
		(1.33%)	(6.67%)	(98.67%)	(93.33%)

$$\text{อัตราการจำแนกกลุ่มผิดพลาดของวิธี DA} = \frac{(7\times100)}{300} = 2.33\%$$

$$\text{อัตราการพยากรณ์ผิดพลาดของวิธี LA} = \frac{(17\times100)}{300} = 5.67\%$$

แสดงว่าการวิเคราะห์จำแนกกลุ่มนี้สามารถสามารถจำแนกกลุ่มผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 97.67% ส่วนการวิเคราะห์การถดถอยโลจิสติกสามารถใช้ในการพยากรณ์กลุ่มผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 94.33%

สำหรับข้อมูลในงานวิจัยนี้ เมื่อพิจารณาจากอัตราความผิดพลาดจากการจำแนกกลุ่มผู้ป่วยออกเป็น 2 กลุ่ม พบว่า วิธีการวิเคราะห์จำแนกกลุ่มมีอัตราความผิดพลาดจากการจำแนกกลุ่มผู้ป่วยน้อยกว่าวิธีการวิเคราะห์การถดถอยโลจิสติก หรืออีกนัยหนึ่งสามารถกล่าวได้ว่าวิธีการวิเคราะห์จำแนกกลุ่มมีประสิทธิภาพสูงกว่าวิธีการวิเคราะห์การถดถอยโลจิสติก

สรุปและข้อเสนอแนะ

สรุป

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการวินิจฉัยการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโรงพยาบาลเมืองฉะเชิงเทราที่ได้รับการตรวจการทำงานของตับระหว่างเดือนกันยายน 2548 – ตุลาคม 2549 ซึ่งมีจำนวนทั้งสิ้นประมาณ 1200 ราย และกลุ่มงานเทคนิคการแพทย์โรงพยาบาลเมืองฉะเชิงเทราได้ทำการจำแนกกลุ่มผู้ป่วยโรคตับออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และกลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ ซึ่งผู้วิจัยได้สุ่มตัวอย่างผู้ป่วยทั้ง 2 กลุ่มมาศึกษากลุ่มละ 150 ราย ซึ่งในงานวิจัยนี้ได้ศึกษาเปรียบเทียบเทคนิคการวิเคราะห์ทางสถิติ 2 วิธี คือ การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) และการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) โดยพิจารณาจากค่าอัตราการพยากรณ์ที่ผิดพลาด (Apparent Error Rate: APER) โดยวิธีที่มีประสิทธิภาพสูงกว่าเป็นวิธีที่มีค่าอัตราการพยากรณ์ที่ผิดพลาดต่ำสุด เพื่อให้ได้ตัวแบบที่ใช้ในการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วย ซึ่งตัวแปรอิสระที่ทำการศึกษาในงานวิจัยนี้มี 8 ตัวแปร คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) และจากการวิจัยดังกล่าวข้างต้นสามารถสรุปผลได้ดังนี้

1. การวิเคราะห์จำแนกกลุ่ม

ข้อมูลผู้ป่วยที่ได้นำมาทำการศึกษาในครั้งนี้เมื่อทำการตรวจสอบสมมติฐานเบื้องต้น พบว่าไม่เป็นไปตามที่กำหนดกล่าวคือ ความแปรปรวนร่วม ไม่เท่ากัน ดังนั้นจึงไม่เหมาะสมที่จะใช้ฟังก์ชันจำแนกกลุ่มเชิงเส้นตรง (Linear Discriminant Function)ในการพยากรณ์ ผู้วิจัยจึงเลือกใช้ฟังก์ชันจำแนกประเภทกำลังสอง (Quadratic Discriminant Function)ในการสร้างตัวแบบพยากรณ์กลุ่มผู้ป่วยแทนเพื่อเป็นการแก้ไขปัญหาข้างต้น

โดยมีฟังก์ชันจำแนกประเภทกำลังสอง (Quadratic Discriminant Function) ดังต่อไปนี้

กลุ่มที่ 1 (กลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ)

$$\begin{aligned}\hat{Y}_1 = {} & -109.090 + 12.996x_1 + 11.447x_2 + 12.969x_3 + 16.128x_4 - 4.460x_5 \\ & + 0.238x_6 - 0.015x_7 + 0.245x_8 + 105.718x_1x_2 + 105.896x_1x_3 \\ & + 0.158x_1x_4 - 1.152x_1x_5 + 0.114x_1x_6 - 0.052x_1x_7 - 0.018x_1x_8 \\ & - 103.452x_2x_3 - 0.350x_2x_4 - 3.432x_2x_5 - 0.154x_2x_6 + 0.102x_2x_7 \\ & - 0.002x_2x_8 - 1.702x_3x_4 + 6.850x_3x_5 - 0.050x_3x_6 + 0.002x_3x_7 \\ & + 0.018x_3x_8 + 13.890x_4x_5 + 0.048x_4x_6 + 0.018x_4x_7 - 0.040x_4x_8 \\ & - 0.158\, x_5x_6 + 0.006x_5x_7 + 0.126x_5x_8 + 0.006x_6x_7 - 53.741x_1^2 \\ & - 54.977x_2^2 - 56.769x_3^2 - 8.797x_4^2 - 51.671x_5^2 - 0.008x_6^2 - 0.005x_7^2 \\ & - 0.001x_8^2\end{aligned}$$

กลุ่มที่ 2 (กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ)

$$\begin{aligned}\hat{Y}_2 = {} & -27.429 - 1.910x_1 + 6.803x_2 + 6.263x_3 + 1.581x_4 - 0.866x_5 \\ & + 0.018x_6 - 0.006x_7 + 0.006x_8 + 112.122x_1x_2 + 112.296x_1x_3 \\ & + 0.608x_1x_4 - 0.778x_1x_5 + 0.012x_1x_6 - 0.002x_1x_7 + 0.002x_1x_8 \\ & - 112.030x_2x_3 - 0.306x_2x_4 + 0.060x_2x_5 - 0.014x_2x_6 + 0.006x_2x_7 \\ & - 0.002x_2x_8 - 0.484x_3x_4 + 0.612x_3x_5 - 0.006x_3x_6 - 0.002x_3x_7 \\ & + 0.002x_3x_8 + 3.392x_4x_5 - 0.006x_4x_6 + 0.002x_4x_7 - 0.004x_4x_8 \\ & + 0.014x_5x_6 - 0.004x_5x_7 + 0.006x_5x_8 - 56.158x_1^2 - 56.767x_2^2 \\ & - 56.947x_3^2 - 1.413x_4^2 - 2.460x_5^2\end{aligned}$$

เมื่อพิจารณาความเหมาะสมของฟังก์ชันจำแนกกลุ่มกำลังสอง พบว่า มีความคลาดเคลื่อนจากการพยากรณ์ หรือจำแนกกลุ่มประมาณ 2.33% หรือกล่าวได้ว่าฟังก์ชันจำแนกกลุ่มกำลังสองนี้สามารถใช้ในการพยากรณ์ หรือจำแนกกลุ่มผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 97.67%

2. ผลการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis)

ข้อมูลผู้ป่วยที่ได้นำมาทำการศึกษาในครั้งนี้เมื่อทำการตรวจสอบสมมติฐานเบื้องต้น พบว่าไม่เป็นไปตามที่กำหนดกล่าวคือ ตัวแปรอิสระมีความสัมพันธ์กัน หรือเกิดปัญหา Multicollinearity ผู้วิจัยจึงทำการการรวมตัวแปรอิสระโดยมีการปรับขนาดของตัวแปรอิสระให้มีค่าของตัวแปรเป็นค่ามาตรฐานแล้วจึงนำข้อมูลของตัวแปรมารวมกันเพื่อสร้างเป็นตัวแปรใหม่ ซึ่งเป็นวิธีการที่ใช้ในการแก้ไขปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กัน จากนั้นนำตัวแปรที่มีการปรับค่า และตัวแปรที่เหลือมาวิเคราะห์การถดถอยโลจิสติคโดยใช้ตัวแบบโลจิทต่อไป

ซึ่งมีสมการสำหรับการพยากรณ์ดังต่อไปนี้

$$\hat{g}(x) = 13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8$$

$$P(\text{เกิดเหตุการณ์}) = \frac{e^{13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8}}{1 + e^{13.2738 + 0.1503(x_1 + x_2 + x_3) - 3.7401x_4 + 0.9760x_5 - 0.1154x_6 + 0.00787x_7 - 0.0466x_8}}$$

โดยมีอัตราความผิดพลาดจากการพยากรณ์แนวโน้มการเป็น หรือไม่เป็นโรคตับของผู้ป่วยโดยรวมเท่ากับ 5.67% หรือกล่าวได้ว่าฟังก์ชันการถดถอยนี้สามารถใช้ในการพยากรณ์ผู้ป่วยโรคตับได้ด้วยความถูกต้องแม่นยำประมาณ 94.33%

3. เปรียบเทียบประสิทธิภาพของการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก

เมื่อเปรียบเทียบอัตราความผิดพลาดจากการจำแนกกลุ่มผู้ป่วยออกเป็น 2 กลุ่ม พบว่า วิธีการวิเคราะห์จำแนกกลุ่มมีอัตราความผิดพลาดจากการจำแนกกลุ่มผู้ป่วย 2.33% ส่วนวิธีการวิเคราะห์การถดถอยโลจิสติกมีอัตราความผิดพลาดจากการจำแนกกลุ่มผู้ป่วย 5.67% ดังนั้นถ้าไม่คำนึงถึงความยากง่ายในการคำนวณควรใช้วิธีการวิเคราะห์ด้วยวิธีจำแนกกลุ่มมากกว่าการวิเคราะห์การถดถอยโลจิสติก เนื่องจากวิธีการจำแนกกลุ่มมีประสิทธิภาพสูงกว่าการวิเคราะห์การถดถอยโลจิสติก และการวิเคราะห์การถดถอยโลจิสติกนั้นมีตัวแบบที่ไม่เหมาะสมกับลักษณะข้อมูลในงานวิจัยนี้

4. ตัวอย่างการนำตัวแบบที่สร้างโดยวิธีการวิเคราะห์จำแนกกลุ่มไปใช้ในการพยากรณ์

ผู้วิจัยได้ทำการเก็บข้อมูลผู้ป่วยที่ได้รับการตรวจการทำงานของตับครั้งแรก ณ เดือนธันวาคม 2549 – กุมภาพันธ์ 250 จำนวน 100 ราย คือ ผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ และผู้ป่วยที่มีแนวโน้มเป็นโรคตับกลุ่มละ 50 ราย นำมาพยากรณ์แนวโน้มการเป็นโรคตับโดยใช้ตัวแบบการวิเคราะห์จำแนกกลุ่มที่ได้จากการวิเคราะห์ในงานวิจัยนี้ เพื่อตรวจสอบความถูกต้องของตัวแบบ ซึ่งได้ผลการวิเคราะห์ดังนี้

4.1 ผลการวิเคราะห์ข้อมูลเบื้องต้น

ตารางที่ 17 แสดงค่าสถิติเบื้องต้นของข้อมูลสำหรับตัวแปรอิสระทั้ง 8 ตัว ที่ทำการศึกษาจากหน่วยตัวอย่างใหม่จำนวน 100 ราย

ตัวแปร	ค่าปกติ	ค่าเฉลี่ย (Mean)	ค่ามัธยฐาน (Median)	ค่าฐานนิยม (Mode)	ค่าสูงสุด (Max)	ค่าต่ำสุด (Min)	ส่วนเบี่ยงเบนมาตรฐาน (SD.)
x_1	6 - 8 g%	7.488	7.6	7.6	9.9	5.1	1.0094
x_2	3.4 - 5.5 g%	3.785	4.0	4.1	5.2	1.6	0.7070
x_3	2 - 3.5 g%	3.696	3.6	3.6	6.3	1.3	0.8236
x_4	0.1 - 1.1 mg%	1.391	0.7	0.5	0.1	7.5	2.2599
x_5	0 - 0.5 mg%	0.609	0.1	0.0	11.8	0.0	1.8018
x_6	0 - 45 U/L	69.660	39.0	25.0	376.0	11.0	73.8163
x_7	9 - 72 U/L	50.770	38.0	13.0	247.0	9.0	46.6272
x_8	38 – 126 U/L	137.260	86	57	762	37	134.6201

4.2 ผลการวิเคราะห์จำแนกกลุ่ม

การตรวจสอบการประยุกต์ใช้นี้จะใช้ตัวแบบที่ได้จากฟังก์ชันกำลังสองที่ได้จากการวิเคราะห์จำแนกกลุ่ม โดยมีตัวแบบดังนี้

กลุ่มที่ 1 (กลุ่มผู้ป่วยที่มีแนวโน้มไม่เป็นโรคตับ)

$$\begin{aligned}\hat{Y}_1 = {} & -109.090 + 12.996x_1 + 11.447x_2 + 12.969x_3 + 16.128x_4 - 4.460x_5 \\ & + 0.238x_6 - 0.015x_7 + 0.245x_8 + 105.718x_1x_2 + 105.896x_1x_3 \\ & + 0.158x_1x_4 - 1.152x_1x_5 + 0.114x_1x_6 - 0.052x_1x_7 - 0.018x_1x_8 \\ & - 103.452x_2x_3 - 0.350x_2x_4 - 3.432x_2x_5 - 0.154x_2x_6 + 0.102x_2x_7 \\ & - 0.002x_2x_8 - 1.702x_3x_4 + 6.850x_3x_5 - 0.050x_3x_6 + 0.002x_3x_7 \\ & + 0.018x_3x_8 + 13.890x_4x_5 + 0.048x_4x_6 + 0.018x_4x_7 - 0.040x_4x_8 \\ & - 0.158\ x_5x_6 + 0.006x_5x_7 + 0.126x_5x_8 + 0.006x_6x_7 - 53.741x_1^2 \\ & - 54.977x_2^2 - 56.769x_3^2 - 8.797x_4^2 - 51.671x_5^2 - 0.008x_6^2 - 0.005x_7^2 \\ & - 0.001x_8^2\end{aligned}$$

กลุ่มที่ 2 (กลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคตับ)

$$\begin{aligned}\hat{Y}_2 = {} & -27.429 - 1.910x_1 + 6.803x_2 + 6.263x_3 + 1.581x_4 - 0.866x_5 \\ & + 0.018x_6 - 0.006x_7 + 0.006x_8 + 112.122x_1x_2 + 112.296x_1x_3 \\ & + 0.608x_1x_4 - 0.778x_1x_5 + 0.012x_1x_6 - 0.002x_1x_7 + 0.002x_1x_8 \\ & - 112.030x_2x_3 - 0.306x_2x_4 + 0.060x_2x_5 - 0.014x_2x_6 + 0.006x_2x_7 \\ & - 0.002x_2x_8 - 0.484x_3x_4 + 0.612x_3x_5 - 0.006x_3x_6 - 0.002x_3x_7 \\ & + 0.002x_3x_8 + 3.392x_4x_5 - 0.006x_4x_6 + 0.002x_4x_7 - 0.004x_4x_8 \\ & + 0.014x_5x_6 - 0.004x_5x_7 + 0.006x_5x_8 - 56.158x_1^2 - 56.767x_2^2 \\ & - 56.947x_3^2 - 1.413x_4^2 - 2.460x_5^2\end{aligned}$$

โดยที่

x_1 คือ Total Protein (TP)

x_2 คือ Albumin (Alb)

x_3 คือ Globulin (Glb)

x_4 คือ Total Bilirubin (TB)

x_5 คือ Direct Bilirubin (DB)

x_6 คือ SGOT

x_7 คือ SGPT

x_8 คือ Alkaline Phosphatase (ALP)

สำหรับการพยากรณ์โดยฟังก์ชันจำแนกกลุ่มกำลังสองของหน่วยตัวอย่างใหม่นั้นทำได้โดยแทนค่าตัวแปรทั้ง 8 ลงในสมการ $\hat{Y}_1$ และ $\hat{Y}_2$ โดยมีเกณฑ์ในการพิจารณาเพื่อจัดกลุ่มผู้ป่วยดังนี้

ถ้า $\hat{Y}_1 > \hat{Y}_2$ จะพยากรณ์ให้ผู้ป่วยอยู่ในกลุ่มที่ 1 (ผู้ป่วยมีแนวโน้มไม่เป็นโรคตับ)
ถ้า $\hat{Y}_1 < \hat{Y}_2$ จะพยากรณ์ให้ผู้ป่วยอยู่ในกลุ่มที่ 2 (ผู้ป่วยมีแนวโน้มเป็นโรคตับ)

เมื่อแทนค่าหน่วยตัวอย่างใหม่ในตัวแบบแล้วจะได้ผลการจำแนกกลุ่มของแต่ละหน่วยตัวอย่างใหม่ดังต่อไปนี้

ตารางที่ 18 แสดงผลการจำแนกกลุ่มของหน่วยตัวอย่างใหม่

ลำดับที่	ผลการจำแนกกลุ่มจากหน่วยงานเทคนิคการแพทย์	$\hat{Y}_1$	$\hat{Y}_2$	ผลการจำแนกกลุ่มโดยใช้ฟังก์ชันกำลังสอง
1	ไม่เป็น	-4.87869	-12.2528	ไม่เป็น
2	ไม่เป็น	-18.1631	-14.7949	เป็น
3	ไม่เป็น	-6.38116	-11.8836	ไม่เป็น
4	ไม่เป็น	-3.69119	-11.8172	ไม่เป็น
5	ไม่เป็น	-6.73299	-12.0096	ไม่เป็น
6	ไม่เป็น	-8.67107	-12.7821	ไม่เป็น
7	ไม่เป็น	-4.00709	-10.5552	ไม่เป็น
8	ไม่เป็น	-3.54816	-12.5764	ไม่เป็น
9	ไม่เป็น	-6.9072	-12.4984	ไม่เป็น
10	ไม่เป็น	-6.63075	-11.7315	ไม่เป็น
11	ไม่เป็น	-4.66425	-10.3023	ไม่เป็น
12	ไม่เป็น	-5.40814	-10.9586	ไม่เป็น
13	ไม่เป็น	-7.21641	-12.6053	ไม่เป็น
14	ไม่เป็น	-2.91419	-11.4461	ไม่เป็น
15	ไม่เป็น	-14.9356	-12.5433	เป็น
16	ไม่เป็น	-4.66425	-10.3023	ไม่เป็น
17	ไม่เป็น	-12.9063	-15.915	ไม่เป็น
18	ไม่เป็น	-12.7109	-13.3203	ไม่เป็น
19	ไม่เป็น	-8.64679	-10.4721	ไม่เป็น
20	ไม่เป็น	-6.99108	-12.4155	ไม่เป็น
21	ไม่เป็น	-3.09746	-12.3403	ไม่เป็น

ตารางที่ 18 (ต่อ)

ลำดับที่	ผลการจำแนกกลุ่มจากหน่วยงานเทคนิคการแพทย์	$\hat{Y}_1$	$\hat{Y}_2$	ผลการจำแนกกลุ่มโดยใช้ฟังก์ชันกำลังสอง
22	ไม่เป็น	-3.97054	-11.4498	ไม่เป็น
23	ไม่เป็น	-7.48411	-12.2224	ไม่เป็น
24	ไม่เป็น	-12.2341	-13.7633	ไม่เป็น
25	ไม่เป็น	-9.0737	-12.5494	ไม่เป็น
26	ไม่เป็น	-11.1503	-12.9595	ไม่เป็น
27	ไม่เป็น	-6.35301	-11.7639	ไม่เป็น
28	ไม่เป็น	-6.58625	-11.3914	ไม่เป็น
29	ไม่เป็น	-8.55617	-11.8542	ไม่เป็น
30	ไม่เป็น	-6.31684	-12.4556	ไม่เป็น
31	ไม่เป็น	-15.5932	-16.2865	ไม่เป็น
32	ไม่เป็น	-5.58355	-11.6759	ไม่เป็น
33	ไม่เป็น	-12.0151	-14.2104	ไม่เป็น
34	ไม่เป็น	-9.43352	-11.3219	ไม่เป็น
35	ไม่เป็น	-10.5838	-11.7135	ไม่เป็น
36	ไม่เป็น	-4.94931	-11.7986	ไม่เป็น
37	ไม่เป็น	-9.18066	-11.9523	ไม่เป็น
38	ไม่เป็น	-9.43449	-12.7218	ไม่เป็น
39	ไม่เป็น	-7.35758	-11.2209	ไม่เป็น
40	ไม่เป็น	-8.8634	-11.2922	ไม่เป็น
41	ไม่เป็น	-4.46864	-11.6928	ไม่เป็น
42	ไม่เป็น	-4.38566	-11.1098	ไม่เป็น
43	ไม่เป็น	-3.52454	-10.5322	ไม่เป็น
44	ไม่เป็น	-4.66085	-10.7204	ไม่เป็น
45	ไม่เป็น	-4.36299	-10.5099	ไม่เป็น
46	ไม่เป็น	-4.91276	-9.90016	ไม่เป็น
47	ไม่เป็น	-6.07309	-11.2893	ไม่เป็น
48	ไม่เป็น	-7.79352	-9.85767	ไม่เป็น
49	ไม่เป็น	-4.29929	-11.3958	ไม่เป็น
50	ไม่เป็น	-4.87869	-12.2528	ไม่เป็น
51	เป็น	-73.411	-15.5855	เป็น
52	เป็น	-399.477	-7.673	เป็น

ตารางที่ 18 (ต่อ)

ลำดับที่	ผลการจำแนกกลุ่มจากหน่วยงานเทคนิคการแพทย์	$\hat{Y}_1$	$\hat{Y}_2$	ผลการจำแนกกลุ่มโดยใช้ฟังก์ชันกำลังสอง
53	เป็น	-549.517	2.01979	เป็น
54	เป็น	-81.368	-6.99186	เป็น
55	เป็น	-586.425	-14.0152	เป็น
56	เป็น	-37.9825	-9.88299	เป็น
57	เป็น	-82.3927	-7.86056	เป็น
58	เป็น	-146.245	-10.7314	เป็น
59	เป็น	-30.5026	-10.4117	เป็น
60	เป็น	-187.589	-3.66071	เป็น
61	เป็น	-2026.69	-5.74969	เป็น
62	เป็น	-32.746	-11.5325	เป็น
63	เป็น	-5825.84	-15.2499	เป็น
64	เป็น	-34.6419	-9.92468	เป็น
65	เป็น	-32.7271	-10.2847	เป็น
66	เป็น	-6679.76	-51.425	เป็น
67	เป็น	-44.8091	-9.22	เป็น
68	เป็น	-29.1012	-12.0174	เป็น
69	เป็น	-24.105	-8.78988	เป็น
70	เป็น	-173.585	-7.25844	เป็น
71	เป็น	-40.0013	-10.3577	เป็น
72	เป็น	-16.1382	-9.80819	เป็น
73	เป็น	-25.7105	-9.36736	เป็น
74	เป็น	-144.778	-6.60388	เป็น
75	เป็น	-19.9796	-8.6109	เป็น
76	เป็น	-124.84	-6.06831	เป็น
77	เป็น	-188.751	-1.74812	เป็น
78	เป็น	-183.564	-3.3242	เป็น
79	เป็น	-29.3346	-10.7675	เป็น
80	เป็น	-14.9406	-9.34855	เป็น
81	เป็น	-4.11654	-11.7849	ไม่เป็น
82	เป็น	-35.1398	-8.82526	เป็น
83	เป็น	-252.984	-4.2434	เป็น

ตารางที่ 18 (ต่อ)

ลำดับที่	ผลการจำแนกกลุ่มจากหน่วยงานเทคนิคการแพทย์	$\hat{Y}_1$	$\hat{Y}_2$	ผลการจำแนกกลุ่มโดยใช้ฟังก์ชันกำลังสอง
84	เป็น	-45.4947	-7.79943	เป็น
85	เป็น	-308.037	1.82368	เป็น
86	เป็น	-30.3374	-9.4349	เป็น
87	เป็น	-53.1816	-8.14356	เป็น
88	เป็น	-471.732	4.25499	เป็น
89	เป็น	-31.0629	-11.1856	เป็น
90	เป็น	-39.4979	-10.0304	เป็น
91	เป็น	-116.285	-9.5189	เป็น
92	เป็น	-87.6788	-12.1901	เป็น
93	เป็น	-55.5354	-13.7307	เป็น
94	เป็น	-885.11	-1.08155	เป็น
95	เป็น	-76.0603	-13.3233	เป็น
96	เป็น	-590.999	-1.86548	เป็น
97	เป็น	-127.417	-14.3125	เป็น
98	เป็น	-90.121	-17.3324	เป็น
99	เป็น	-417.073	-10.3007	เป็น
100	เป็น	-2808.69	-14.6602	เป็น

และสามารถสรุปผลการจำแนกกลุ่มของหน่วยตัวอย่างใหม่ดังตารางต่อไปนี้

ตารางที่ 19 แสดงผลการประเมินการจัดกลุ่มของหน่วยตัวอย่างใหม่โดยใช้ตัวแบบที่ได้จากวิธีการวิเคราะห์จำแนกกลุ่ม

กลุ่ม	ขนาดตัวอย่าง (n_i)	ผลการจำแนกกลุ่ม	
		จำแนกกลุ่มถูก	จำแนกกลุ่มผิด
มีแนวโน้มไม่เป็นโรคตับ	50	48	2
		(9600%)	(4%)
มีแนวโน้มเป็นโรคตับ	50	49	1
		(98.00%)	(2%)

ดังนั้น อัตราการจำแนกกลุ่มผิดพลาด = $\frac{(3\times100)}{100}$ = 3.00%

เมื่อพิจารณาผลการจัดกลุ่มหน่วยตัวอย่างใหม่ของผู้ป่วยที่มีแนวโน้มเป็นโรคตับ และไม่มีแนวโน้มเป็นโรคตับจำนวน 100 ราย พบว่า มีการจำแนกลุ่มผู้ป่วยผิดพลาดจำนวน 3 ราย นั่นคือพยากรณ์ผิดพลาด 3.00% หรือกล่าวอีกนัยหนึ่งได้ว่าตัวแบบนี้สามารถจำแนกกลุ่มผู้ป่วยโรคตับได้ถูกต้องแม่นยำประมาณ 97.00%

ข้อเสนอแนะ

1. จากการตรวจสอบความเหมาะสมของตัวแบบในการวิเคราะห์การถดถอยโลจิสติก พบว่าตัวแบบที่สร้างไม่เหมาะสมกับลักษณะข้อมูลที่นำมาศึกษาในงานวิจัยนี้ ดังนั้นในการศึกษาครั้งต่อไปอาจจะใช้ตัวแบบโพรบิทในการวิเคราะห์ ซึ่งอาจจะให้ค่าที่เหมาะสมมากกว่า เนื่องจากตัวแบบโพรบิทสามารถนำมาใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ เมื่อตัวแปรตามเป็นตัวแปรที่มี 2 ลักษณะ คือ มีค่าเท่ากับ 1 และ 0 ได้เช่นกัน ซึ่งการวิเคราะห์ตัวแบบโพรบิทเป็นการใช้ฟังก์ชันการแจกแจงสะสมปกติมาตรฐาน (Standard normal cumulative distribution function) แปลงค่าเฉลี่ยของตัวแปลตาม (Z_i) ให้เป็นความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (P_i)

2. งานวิจัยนี้วิเคราะห์โดยใช้ตัวแปรอิสระเพียง 8 ตัวเท่านั้น คือ Total protein (TP: x_1), Albumin (Alb: x_2), Globulin (Glb: x_3), Total Bilirubin (TB: x_4), Direct Bilirubin (DB: x_5), SGOT (x_6), SGPT (x_7) และAlkaline Phosphatase (ALP: x_8) ดังนั้นในการศึกษาครั้งต่อไปผู้วิจัยอาจเพิ่มตัวแปรอิสระที่เกี่ยวข้องในการพยากรณ์ผู้ป่วยโรคตับ ได้แก่ อาการตัวเหลือง, ตาเหลือง เป็นต้น เพื่อช่วยเพิ่มประสิทธิภาพในการพยากรณ์

เอกสารและสิ่งอ้างอิง

กนกศรี ศรินนภากร. 2540. **การรู้จำลายมือเขียนตัวอักษรภาษาไทยโดยวิธีการจำแนกกลุ่ม**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

กัลยา วานิชย์บัญชา. 2544. **การวิเคราะห์ตัวแปรหลายตัวด้วย SPSS forWindows**. โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

กัลยา วานิชย์บัญชา. 2548. **การวิเคราะห์สถิติขั้นสูงด้วย SPSS forWindows**. พิมพ์ครั้งที่ 4. โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ชนิศวรา ฉัตรแก้ว. 2543. **การวิเคราะห์การถดถอยเมื่อตัวแปรตามมีสองลักษณะโดยใช้ตัวแบบความน่าจะเป็นเชิงเส้น ตัวแบบโพรบิท และตัวแบบโลจิท**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาเกษตรศาสตร์.

ชมรมโรคตับแห่งประเทศไทย. **การตรวจการทำงานของตับ.** แหล่งที่มา: http://thailiverclub.org/liverfunc.htm, 16 กรกฎาคม 2549.

มาลินี ศรีคำม้วน. 2548. **รู้จริง รู้ชัด ไวรัสตับอักเสบ**. พิมพ์ครั้งที่ 1. สำนักพิมพ์ใกล้หมอ, กรุงเทพฯ.

ยี่สุ่น นันทวโนทยาน. 2539. **การวิเคราะห์ตัวแปรจำแนกกลุ่มนักศึกษาที่เลือกและไม่เลือกเรียนวิชาชีพการพยาบาล เขตกรุงเทพมหานคร**. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

สรินยา สำเภาเงิน. 2542. **การเปรียบเทียบการวิเคราะห์จำแนกประเภท กับการวิเคราะห์การถดถอยมัลติโนเมียลในการศึกษาปัจจัยที่มีผลต่อระดับคะแนนของนิสิตมหาวิทยาลัยเกษตรศาสตร์ปีเข้าศึกษา 2538**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาเกษตรศาสตร์.

สุชาต ประสิทธิ์รัฐสินธุ์. 2548. **เทคนิคการวิเคราะห์ตัวแปรหลายตัวสำหรับการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์**. พิมพ์ครั้งที่ 5. หจก.สามลดา, กรุงเทพฯ.

สุรเชษฐ ขันธิโชติ. 2547. **การวิเคราะห์จำแนกปัจจัยที่ส่งผลกระทบต่อการชำระหนี้คืนของเกษตรกรที่เป็นสมาชิกในศูนย์พัฒนาโครงการหลวงห้วยลึก จังหวัดเชียงใหม่**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยชียงใหม่.

สุวรรณา พูนกล้า. 2538. **สมการจำแนกกลุ่มการเรียนวิชาชีพ โดยใช้ความถนัดทาง การเรียนเป็น ตัวจำแนก สำหรับนักเรียนช่างอุตสาหกรรม ระดับประกาศนียบัตรวิชาชีพ**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยชียงใหม่.

อรัญญา ศรียัพ. 2537. **การเปรียบเทียบการวิเคราะห์การถดถอยแบบโลจิสติก กับการวิเคราะห์จำแนกประเภทในการศึกษาปัจจัยที่มีผลต่อความสำเร็จในการศึกษาของนักเรียนนายทหารโรงเรียนนายเรือ**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

อุไรวรรณ อมรนิมิตร. 2546. **เทคนิคการวิเคราะห์ความเสี่ยง : ศึกษาเปรียบเทียบระหว่าง Logistic Regression Analysis และ Discriminant Analysis**. วารสารสถิติประยุกต์: 60 – 66.

Barbara, G. and S. Linda. 2007. **Using Multivariate Statistics**. 5th ed. Pearson Education, Inc., Boston.

Daxin, J., T. Chun and Z. Aidong. n.d. **Cluster Analysis for Gene Expression Data : A Survey.** Available Source: http://www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/survey.pdf, May 15, 2006.

Efron, B. 1975. The Efficiency of logistic regression compared to normal discriminant analysis. **J. Amer. Statist. Ass.** 70: 892 – 898.

Johnson, A. 2002. **Applied Multivariate Statistical Analysis**. 5th ed. Prentice-Hall, Inc., New Jersey.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lutkepohl and T.C. Lee. 1998. **Introduction to the Theory ND Practice of Econometrics**. 2nd ed. John Wiley and Sons, Inc., New York.

Press, S.T. and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. **J. Amer. Statist. Ass. 75: 699 – 705.**

Ravinda, K. and N. Dayanand. 2000. **Multivariate Data Reduction and Discriminant with SAS Software**. John Wiley and Sons, Inc., New York.

Rencher and C. Alvin. **Methods of multivariate analysis**. New York : John Wiley & Sons, Inc., 1995.

Worthington, L.H. and C.W. Grant. 1971. Factors of academic success : A multivariate analysis. **The J. Educ. Res.** 65: 7 – 10.

ภาคผนวก

ภาคผนวก ก

ตัวอย่างโปรแกรมสำหรับการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก
โดยโปรแกรม SAS

โปรแกรมที่ใช้สำหรับการทดสอบสมมติฐาน และวิเคราะห์สำหรับวิธีการจำแนกกลุ่ม และวิธีการถดถอยโลจิสติก มีรายละเอียดดังต่อไปนี้

1. การวิเคราะห์การจำแนกกลุ่ม

1.1 โปรแกรมการตรวจสอบสมมติฐาน: การเท่ากันของเมทริกซ์แปรปรวนร่วม

```
data liver;
infile 'liver.dat';
input results x1 x2 x3 x4 x5 x6 x7 x8;
    proc discrim data = liver method = normal pool = test  wcov pcov bcov;
    class results ;
    var x1 x2 x3 x4 x5 x6 x7 x8 ;
run;
```

1.2 โปรแกรมการวิเคราะห์การจำแนกกลุ่ม

```
data liver;
infile 'liver.dat';
input results x1 x2 x3 x4 x5 x6 x7 x8;
    proc discrim data=liver outstat=liverstat wcov pcov method=normal pool=no
    distance anova listerr crosslisterr;
    class results;
    var x1 x2 x3 x4 x5 x6 x7 x8;
run;
```

2. การวิเคราะห์การถดถอยโลจิสติก

2.1 โปรแกรมการตรวจสอบสมมติฐาน : ตัวแปรอิสระไม่ควรมีความสัมพันธ์กัน

```
data liver;
infile ‘liver.dat’;
input results x1 x2 x3 x4 x5 x6 x7 x8;
    proc reg;
    model results = x1 x2 x3 x4 x5 x6 x7 x8 /dw vif;
run;
    proc standard data = liver mean = 0 std = 1 vardef = n out = zscores;
    var x1 x2 x3;
run;
    z123 = x1+x2+x3;
    proc reg;
    model results = z123 x4 x5 x6 x7 x8 /dw vif;
run;
```

2.2 โปรแกรมการวิเคราะห์การถดถอยโลจิสติก

```
data liver;
infile ‘liver.dat’;
input results z123 x4 x5 x6 x7 x8;
    proc logistic data = liver descending outest = est covout nosimple ;
    model results = z123 x4 x5 x6 x7 x8 /risklimits link=logit covb ctable pprob=.5 lackfit ;
    output out=predict predprobs = cross p = phat lower = lcl upper = ucl;
run;
```

ภาคผนวก ข

ผลการวิเคราะห์จำแนกกลุ่ม และการวิเคราะห์การถดถอยโลจิสติก

ผลการทดสอบสมมติฐาน และผลวิเคราะห์วิธีการจำแนกกลุ่ม และการถดถอยโลจิสติกดังนี้

1. ผลการวิเคราะห์การจำแนกกลุ่ม

1.1 ผลการตรวจสอบสมมติฐาน : การเท่ากันของเมทริกซ์แปรปรวนร่วม

The DISCRIM Procedure

Observations	300	DF Total	299
Variables	8	DF Within Classes	298
Classes	2	DF Between Classes	1

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
1591.948454	36	<.0001

1.2 ผลการวิเคราะห์ฟังก์ชันจำแนกกลุ่มกำลังสอง

The DISCRIM Procedure

Observations	300	DF Total	299
Variables	8	DF Within Classes	298
Classes	2	DF Between Classes	1

Class Level Information

Variable results	Name	Frequency	Weight	Proportion	Prior Probability
1	_1	150	150.0000	0.500000	0.500000
2	_2	150	150.0000	0.500000	0.500000

Within Covariance Matrix Information results	Natural Log of theCovariance Matrix Rank	Determinant of the Covariance Matrix
1	8	18.57933
2	8	-0.57125

Multivariate Statistics and Exact F Statistics

S=1 M=3 N=144.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Hotelling-Lawley Trace	1.43227187	52.10	8	291	<.0001

Cross-validation Results using Quadratic Discriminant Function Classified

Obs	results	results	1	2
3	2	1 *	1.0000	0.0000
4	2	1 *	1.0000	0.0000
29	2	1 *	0.8713	0.1287
79	2	1 *	1.0000	0.0000
119	2	1 *	1.0000	0.0000
156	1	2 *	0.0000	1.0000
176	1	2 *	0.0001	0.9999

* Misclassified observation

Number of Observations and Percent Classified into results

From results	1	2	Total
1	148	2	150
	98.67	1.33	100.00
2	5	145	150
	3.33	96.67	100.00
Total	153	147	300
	51.00	49.00	100.00
Priors	0.5	0.5	

Error Count Estimates for results

	1	2	Total
Rate	0.0133	0.0333	0.0233
Priors	0.5000	0.5000	

ตารางผนวกที่ ข1 แสดงค่าสัมประสิทธิ์ของตัวแปรอิสระในการจำแนกกลุ่มที่ 1

ตัวแปร	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Const	-109.090	-109.090	-109.090	-109.090	-109.090	-109.090	-109.090	-109.090
Linear	12.996	11.447	12.969	16.128	-4.460	0.238	-0.015	0.245
x_1	-53.741	52.859	52.948	0.079	-0.576	0.057	-0.026	-0.009
x_2	52.859	-54.977	-51.726	-0.175	-1.716	-0.077	0.051	-0.001
x_3	52.948	-51.726	-56.769	-0.851	3.425	-0.025	0.001	0.009
x_4	0.079	-0.175	-0.851	-8.797	6.945	0.024	0.009	-0.020
x_5	-0.576	-1.716	3.425	6.945	-51.671	-0.079	0.003	0.063
x_6	0.057	-0.077	-0.025	0.024	-0.079	-0.008	0.003	0.000
x_7	-0.026	0.051	0.001	0.009	0.003	0.003	-0.005	0.000
x_8	-0.009	-0.001	0.009	-0.020	0.063	0.000	0.000	-0.001

ตารางผนวกที่ ข2 แสดงค่าสัมประสิทธิ์ของตัวแปรอิสระในการจำแนกกลุ่มที่ 2

ตัวแปร	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Const	-27.429	-27.429	-27.429	-27.429	-27.429	-27.429	-27.429	-27.429
Linear	-1.910	6.803	6.263	1.581	-0.866	0.018	-0.006	0.006
x_1	-56.158	56.061	56.148	0.304	-0.389	0.006	-0.001	0.001
x_2	56.061	-56.767	-56.015	-0.153	0.030	-0.007	0.003	-0.001
x_3	56.148	-56.015	-56.947	-0.242	0.306	-0.003	-0.001	0.001
x_4	0.304	-0.153	-0.242	-1.413	1.696	-0.003	0.001	-0.002
x_5	-0.389	0.030	0.306	1.696	-2.460	0.007	-0.002	0.003
x_6	0.006	-0.007	-0.003	-0.003	0.007	-0.000	0.000	-0.000
x_7	-0.001	0.003	-0.001	0.001	-0.002	0.000	-0.000	0.000
x_8	0.001	-0.001	0.001	-0.002	0.003	-0.000	0.000	-0.000

2. ผลการวิเคราะห์การถดถอยโลจิสติก

2.1 ผลการตรวจสอบสมมติฐาน: ตัวแปรอิสระไม่ควรมีความสัมพันธ์กัน

The REG Procedure

Model: MODEL1

Dependent Variable: results

Number of Observations Read 300

Number of Observations Used 300

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.21754	0.04800	46.20	<.0001
z123	1	0.03594	0.00908	3.96	<.0001
x4	1	-0.23987	0.04116	-5.83	<.0001
x5	1	0.30450	0.05437	5.60	<.0001
x6	1	-0.00491	0.00055608	-8.82	<.0001
x7	1	0.00038278	0.00072572	0.53	0.5983
x8	1	-0.00230	0.00022435	-10.25	<.0001

Parameter Estimates

Variable	DF	Variance Inflation
Intercept	1	0
z123	1	1.06613
x4	1	5.93899
x5	1	6.39674
x6	1	2.28865
x7	1	1.98597
x8	1	1.10830

2.2 ผลการวิเคราะห์ Logistic Regression Analysis

Logistic Regression

The LOGISTIC Procedure

Model Information

Data Set	WORK.LIVER
Response Variable	results
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	300
Number of Observations Used	300

Response Profile

Ordered Value	results	Total Frequency
1	2	150
2	1	150

Probability modeled is results=2.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	417.888	100.496
SC	421.592	126.422
-2 Log L	415.888	86.496

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	329.3924	6	<.0001
Score	161.1249	6	<.0001
Wald	53.9230	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	13.2738	1.9486	46.4024	<.0001
z123	1	0.1503	0.1279	1.3801	0.2401
x4	1	-3.7401	0.9523	15.4245	<.0001
x5	1	0.9760	1.6726	0.3405	0.5595
x6	1	-0.1154	0.0194	35.3364	<.0001
x7	1	0.00787	0.0173	0.2073	0.6489
x8	1	-0.0466	0.00851	29.9681	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
z123	1.162	0.904	1.493
x4	0.024	0.004	0.154
x5	2.654	0.100	70.400
x6	0.891	0.858	0.926
x7	1.008	0.974	1.043
x8	0.954	0.939	0.971

Association of Predicted Probabilities and Observed Responses

Percent Concordant	98.1	Somers' D	0.962
Percent Discordant	1.8	Gamma	0.963
Percent Tied	0.1	Tau-a	0.483
Pairs	22500	c	0.981

Wald Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
z123	1.0000	1.162	0.904	1.493
x4	1.0000	0.024	0.004	0.154
x5	1.0000	2.654	0.100	70.400
x6	1.0000	0.891	0.858	0.926
x7	1.0000	1.008	0.974	1.043
x8	1.0000	0.954	0.939	0.971

Estimated Covariance Matrix

Parameter	Intercept	z123	x4	x5
Intercept	3.797077	0.001006	-1.34328	0.830536
z123	0.001006	0.016362	0.000887	0.015612
x4	-1.34328	0.000887	0.90689	-0.84177
x5	0.830536	0.015612	-0.84177	2.797536
x6	-0.02241	0.000081	0.006362	-0.00389
x7	-0.00794	-0.00017	0.002405	-0.00134
x8	-0.01361	0.000051	0.003648	-0.00313

Estimated Covariance Matrix

Parameter	x6	x7	x8
Intercept	-0.02241	-0.00794	-0.01361
z123	0.000081	-0.00017	0.000051
x4	0.006362	0.002405	0.003648
x5	-0.00389	-0.00134	-0.00313
x6	0.000377	-0.00014	0.000059
x7	-0.00014	0.000299	0.000014
x8	0.000059	0.000014	0.000072

Partition for the Hosmer and Lemeshow Test

		results = 2		results = 1	
Group	Total	Observed	Expected	Observed	Expected
1	36	0	0.00	36	36.00
2	30	0	0.00	30	30.00
3	30	0	0.12	30	29.88
4	30	0	1.22	30	28.78
5	30	12	10.70	18	19.30
6	30	28	25.82	2	4.18
7	30	29	28.87	1	1.13
8	30	29	29.52	1	0.48
9	30	29	29.79	1	0.21
10	24	23	23.95	1	0.05

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
26.2799	8	0.0009

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi tivity	Speci ficity	False POS	False NEG
0.500	143	140	10	7	94.3	95.3	93.3	6.5	4.8

ประวัติการศึกษา และการทำงาน

ชื่อ – นามสกุล	นางสาวศยามล ลำลองรัตน์
วัน เดือน ปี ที่เกิด	10 ตุลาคม 2523
สถานที่เกิด	อำเภอเมือง จังหวัดจันทบุรี
ประวัติการศึกษา	วท.บ. (คณิตศาสตร์) มหาวิทยาลัยบูรพา
ตำแหน่งหน้าที่การงานปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-