บทที่ 6
การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ด้วยวิธีวิเคราะห์การถดถอยแบบโลจิสติก

เทคนิคการจำแนกกลุ่มข้อมูล มีหลายเทคนิควิธีการได้รับการยอมรับ แต่ละเทคนิควิธีการเหล่านั้นมีมุมมองในทางทฤษฎีที่แตกต่างกันไป จากบทที่ผ่านมา เทคนิคการวิเคราะห์การจำแนกประเภท เป็นวิธีการหนึ่งที่อาศัยค่าความแปรปรวนในแต่ละกลุ่มและระหว่างกลุ่มข้อมูล สำหรับสร้างโมเดลเพื่อจำแนกกลุ่มข้อมูล สำหรับในบทนี้ จะอาศัยวิธีการที่เรียกว่า การวิเคราะห์ถดถอยโลจิสติก ซึ่งเป็น กรณีหนึ่งของการวิเคราะห์ถดถอย ในการประมาณค่าของตัวแปรตอบสนองที่ไม่ทราบค่า โดยค่าตัวแปรดังกล่าวนั้น อยู่ในลักษณะของกลุ่มข้อมูล หรือมี ลักษณะเป็นค่าไม่ต่อเนื่อง(Discrete) ผลก็คือ การวิเคราะห์ถดถอยโลจิสติก ถือเป็นวิธีการหนึ่งที่นำมาใช้ในการจำแนกประเภทของกลุ่มข้อมูลได้ ทั้งนี้วิธีการวิเคราะห์ดังกล่าว มีการใช้กันอย่างกว้างขวาง ซึ่งอาจจะมากกว่าเทคนิคการวิเคราะห์การจำแนกประเภทดังที่กล่าวมาแล้วด้วยซ้ำไป โดยเฉพาะในทางการแพทย์ ตัวอย่างหนึ่งที่เห็นได้ชัดคือ นำไปใช้ในการทำนายการเกิดโรคของคนจากปัจจัยต่างๆที่ศึกษา เหตุผลที่เทคนิควิธีการดังกล่าวมีการนำมาใช้ในวงกว้าง ก็เนื่องจากว่า เทคนิคการวิเคราะห์ดังกล่าว สร้างโมเดลของการวิเคราะห์ข้อมูล โดยใช้ตัวแปรอิสระที่ไม่จำเป็นต้องมีการแจกแจงแบบปกติ (Normal Distribution) อีกทั้งตัวแปรอิสระสามารถเป็นได้ทั้ง ลักษณะต่อเนื่อง(Continuous) หรือไม่ต่อเนื่อง และผลที่ได้จากการวิเคราะห์ แสดงออกมาในลักษณะของความน่าจะเป็น ซึ่งช่วยให้ง่ายต่อการทำความเข้าใจ และตีความหมายได้สมบูรณ์ยิ่งขึ้น

ในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ เพื่อจำแนกประเภทของข้อมูล จะอาศัยข้อมูลชุดเดียวกับที่ใช้ ในบทที่ผ่านมา นั่นคือชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งลิวคีเมีย เพื่อที่จะจำแนกและทำนายกลุ่ม ของข้อมูล โดยใช้ ค่าการแสดงออกของยีนต่างๆ เป็นค่าของตัวแปรอิสระในการสร้างโมเดลจำแนกประเภท นอกจากนี้ เนื่องจากยีนที่เป็นตัวแปรอิสระต่างๆ นั้นมีจำนวนมาก การวิเคราะห์ข้อมูลจำเป็นต้องเลือกตัวแปรอิสระ หรือสร้างตัวแปรอิสระใหม่ที่เป็นตัวแทนของตัวแปรอิสระดังกล่าว สำหรับนำมาเป็นตัวแปรสำหรับจำแนกกลุ่มข้อมูล ซึ่งในที่นี้จะใช้วิธีการเดียวกับวิธีการที่อธิบายในบทที่ 5 ได้แก่ การเลือกตัวแปรโดยวิธีวิเคราะห์ค่าเอนโทรปี การเลือกตัวแปรโดยวิธีการวิเคราะห์ค่าความแปรปรวน และการสร้างตัวแปรใหม่โดยวิธีวิเคราะห์องค์ประกอบหลัก อย่างไรก็ตาม ข้อจำกัดของวิธีการถดถอยโลจิสติกที่ต่างจากวิธีการวิเคราะห์จำแนกประเภท คือ จำนวนของตัวแปรที่ใช้สำหรับ

ทำนายค่า หรือการจำแนกกลุ่มข้อมูล จะต้องมีจำนวนน้อยกว่าจำนวนตัวอย่างข้อมูล ผลก็คือจำนวนตัวแปรที่ใช้วิเคราะห์ในวิธีการนี้ จะมีจำนวนที่น้อยกว่าตัวแปรที่ใช้วิเคราะห์การจำแนกประเภทในบทที่ 5

6.1 หลักการของวิธีวิเคราะห์การถดถอยแบบโลจิสติก

การวิเคราะห์ถดถอยโดยทั่วไปเป็นการวิเคราะห์เพื่อที่จะทำนายค่าของตัวแปรตอบสนอง ซึ่งมีลักษณะเป็นตัวแปรต่อเนื่อง (Continuous Variables) โดยสมการถดถอย แสดงให้เห็นว่า ค่าที่ต้องการทำนายเกิดจากผลรวมเชิงเส้นของตัวแปรอิสระต่างๆ ทั้งนี้ในกรณีของตัวแปรตอบสนองที่มีค่าไม่ใช่ค่าต่อเนื่อง (Discrete Variables) เช่น ตัวแปรที่มี 2 ค่าในลักษณะของการเกิดเหตุการณ์หรือไม่เกิดเหตุการณ์ (Dichotomous Variables) ตัวแปรที่เป็นลักษณะของลำดับ(Ordinary Variables) ตัวแปรกลุ่ม (Category Variables) รูปแบบของสมการถดถอย หรือโมเดลของการถดถอยโดยทั่วไปนั้นไม่สามารถหาคำตอบได้ นักสถิติจึงได้คิดค้นโมเดลของการถดถอยขึ้นมาใหม่ ซึ่งเป็นกรณีเฉพาะของการถดถอย ที่ค่าของตัวแปรตอบสนองที่ต้องการทำนาย มีลักษณะไม่ต่อเนื่อง การวิเคราะห์ถดถอยโดยอาศัยโมเดลดังกล่าวจะเรียกว่า การวิเคราะห์การถดถอยโลจิสติก ซึ่งในงานวิจัยนี้จะใช้วิธีการดังกล่าวในการจำแนกกลุ่มของข้อมูล(Classification) หรือ ทำนายกลุ่มของข้อมูล จากตัวแปรอิสระต่างๆ

จากเมตริกซ์ข้อมูล

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \tag{93}$$

โดยที่ X เป็นเมตริกซ์ข้อมูล มี m เป็นจำนวนของตัวแปรทำนายค่า และ n เป็นจำนวนของตัวอย่างข้อมูล และ y เป็นเวกเตอร์ของตัวแปรที่เป็นกลุ่มข้อมูลที่ต้องการคำนายค่า ซึ่ง

$$y=[y_1, y_2, ..., y_i, ..., y_n]' \tag{94}$$

โมเดลของการถดถอย โดยทั่วไปจะแสดงได้ดังนี้

$$y_i=\beta_0+\beta_1 x_{i1}+...+\beta_j x_{ij}+...+\beta_m x_{im}+\varepsilon_i \tag{95}$$

จากสมการ เมื่อ y เป็นค่าต่อเนื่อง จะสามารถทำนายค่าได้โดยตรงจากสมการ โดยใช้ค่าพารามิเตอร์ β จากการประมาณค่าโดยวิธีกำลังสองน้อยที่สุด(Least squares estimation)

ในกรณีที่ ตัวแปร y ไม่เป็นค่าต่อเนื่องจะไม่สามารถทำนายค่าจากโมเดลดังกล่าวได้โดยตรง การวิเคราะห์ถดถอยโลจิสติก จะใช้หลักการของความน่าจะเป็น (Probability) มาใช้ทำนายค่า y นั่นคือ ค่าที่ได้จากการทำนาย จะอยู่ในรูปของค่าความน่าจะเป็นที่จะเกิด y ซึ่งค่าความน่าจะเป็น จะมีลักษณะของค่าต่อเนื่อง ที่สามารถคำนวณได้จากโมเดล ดังนั้นวิธีการวิเคราะห์ถดถอยโลจิสติก จะต้องรู้ว่าค่า y มีค่าอะไรได้บ้าง (เช่น y เป็นกลุ่มของข้อมูล) การประมาณค่า y จะต้องประมาณค่าความน่าจะเป็นที่ y จะเป็นกลุ่มใดๆ แล้วอาศัยค่าความน่าจะเป็นดังกล่าวในการตัดสินใจว่า y จะอยู่ในกลุ่มใด

จากแนวคิดดังกล่าว ฟังก์ชันเริ่มต้นของการถดถอยโลจิสติก จึงแสดงได้ดังสมการ (96)

$$p(x_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \tag{96}$$

จากสมการ $p(x_i)$ เป็นค่าความน่าจะเป็นที่ y_i จะมีค่าตามที่กำหนดเช่น $y_i = 1$ (ค่าของ y ที่ตัวอย่าง i อยู่ในกลุ่มที่ 1) ดังนั้น $1 - p(x_i)$ จึงเป็นค่าความน่าจะเป็นที่ y_i จะไม่เป็นไปตามข้อกำหนดดังกล่าว

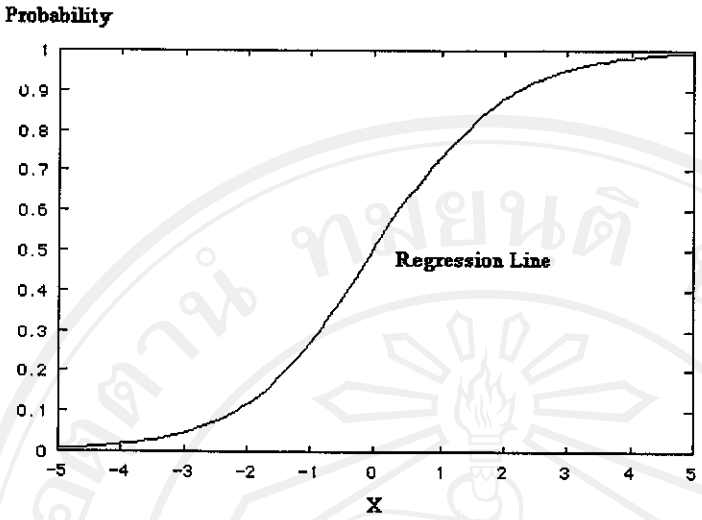
เนื่องจาก $p(x_i)$ เป็นค่าความน่าจะเป็นซึ่งต้องมีค่าอยู่ระหว่าง 0 กับ 1 ดังนั้นเพื่อให้โมเดลของการวิเคราะห์เป็นไปตามข้อกำหนดดังกล่าว สมการตั้งต้นของการวิเคราะห์ถดถอยจึงปรับใหม่เป็น

$$p(x_i) = e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}} \tag{97}$$

จากสมการ (97) ค่า $p(x_i)$ ที่ได้ จะมีค่าอยู่ระหว่าง 0 กับ 1 และเนื่องจากค่าความน่าจะเป็นจะต้องมีค่าเป็นบวกเสมอ ดังนั้นฟังก์ชันการถดถอย จึงปรับใหม่ให้เป็นไปตามเงื่อนไขดังกล่าว และแสดงได้ดังสมการ (98)

$$p(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}} = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij})}}; 0 \le p(x_i) \le 1 \tag{98}$$

ฟังก์ชันที่แสดงในสมการ (98) จะเรียกว่า ฟังก์ชัน โลจิสติก (Logistic Function) ลักษณะของการแจกแจงข้อมูลในฟังก์ชันจากสมการนี้ จะเรียกว่า การแจกแจงแบบโลจิสติก (Logistic Distribution) ซึ่งแสดงได้ดัง รูป 6.1



รูป 6.1 การแจกแจงข้อมูลในลักษณะการแจกแจงแบบโลจิสติก

จากสมการที่ผ่านมาเมื่อพิจารณาสัดส่วนของค่าความน่าจะเป็นที่จะเกิดเหตุการณ์หรือเป็นไปตามข้อกำหนดของตัวแปร y ต่อค่าความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ดังกล่าว จะแสดงได้ดังสมการ (99)

$$\frac{p(x_i)}{1-p(x_i)} = \frac{e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} /(1+e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}})}{1/(1+e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}})} = e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} \qquad (99)$$

สัดส่วนดังกล่าวจะเรียกว่า ออดส์เรโช (Odds Ratio) หรือเขียนย่อๆว่า ค่าโออาร์ (OR) ดังนั้น

$$\ln\left[\frac{p(x_i)}{1-p(x_i)}\right] = \ln(e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}) = \beta_0 + \sum_{j=1}^{m}\beta_j x_{ij} \qquad (100)$$

โมเดลที่ได้จาก สมการ (100) จะเรียกว่า โมเดลของการถดถอยโลจิสติก (Logistic Regression Model) หรือจะเรียกว่า โลจิสโมเดล (Logit Model) โมเดลของการถดถอยโลจิสติกนี้จะนำไปใช้ในการประมาณค่าพารามิเตอร์ β ซึ่งวิธีการที่ใช้ในการประมาณค่าพารามิเตอร์ดังกล่าวเรียกว่า การประมาณค่าพารามิเตอร์ด้วยวิธีการความเป็นไปได้สูงสุด (Maximum Likelihood Estimation: ML)

การประมาณค่าพารามิเตอร์ด้วยวิธีการความเป็นไปได้สูงสุด เป็นวิธีการหาค่าตัวประมาณที่ทำให้ฟังก์ชันความเป็นไปได้ (Likelihood Function) มีค่าสูงที่สุด ในการประมาณค่าพารามิเตอร์

ดังกล่าวจึงเริ่มด้วยการกำหนด ฟังก์ชันล็อกความเป็นไปได้ (log likelihood function) ของโมเดลการถดถอยโลจิสติก ดังสมการ (101)

$$\ln\left[L(p(x))\right]=\sum_{i=1}^{n}\left(y_i\ln\frac{p(x_i)}{1-p(x_i)}+\ln(1-p(x_i))\right) \tag{101}$$

กำหนดให้

$$f(x)=\ln\frac{p(x)}{1-p(x)} \tag{102}$$

ดังนั้น

$$p(x)=\frac{1}{1+e^{-f(x)}} \tag{103}$$

จะได้

$$\ln\left[L(f(x))\right]=\sum_{i=1}^{n}\left(y_i f(x_i)+\ln(1-\frac{1}{1+e^{-f(x_i)}})\right) \tag{104}$$

หรือ

$$\ln\left[L(f(x))\right]=\sum_{i=1}^{n}\left(y_i f(x_i)-\ln(1+e^{f(x_i)})\right) \tag{105}$$

จากสมการ (100) และ สมการ (102) จะได้

$$f(x_i)=\beta_0+\sum_{j=1}^{m}\beta_j x_{ij} \tag{106}$$

เพื่อที่จะประมาณค่าพารามิเตอร์ β ดังนั้น จากฟังก์ชันล็อกความเป็นไปได้ ในสมการ (101) จะได้

$$\ln\left[L(\beta)\right]=\sum_{i=1}^{n}\left(y_i(\beta_0+\sum_{j=1}^{m}\beta_j x_{ij})-\ln(1+e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}})\right) \tag{107}$$

หรือ

$$\ln\left[L(\beta)\right]=\sum_{i=1}^{n}\left(y_i x_i'\beta-\ln(1+e^{x_i'\beta})\right) \tag{108}$$

หาอนุพันธ์ (Differential) อันดับหนึ่งในสมการ (108) เทียบกับพารามิเตอร์ β จะได้

$$\frac{\partial}{\partial\beta}\ln\left[L(\beta)\right]=\sum_{i=1}^{n}\left(y_i x_i'-(1+e^{x_i'\beta})^{-1}e^{x_i'\beta}x_i'\right) \tag{109}$$

นั่นคือ เมื่อกำหนดให้ p เป็นเวกเตอร์ของค่าความน่าจะเป็นที่ค่าของ y จะอยู่ในกลุ่มที่ตั้งสมมุติฐาน จะได้

$$\frac{\partial}{\partial\beta}\ln\left[L(\beta)\right]=X'(y-p) \tag{110}$$

หาอนุพันธ์อันดับสอง ของสมการ (108) เทียบกับพารามิเตอร์ β จะได้

$$\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left[L(\beta)\right]=-\sum_{i=1}^{n}x_i x_i' p(x_i)(1-p(x_i)) \tag{111}$$

หรือ

$$\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left[L(\beta)\right]=-X'WX \tag{112}$$

เมื่อ W คือเมตริกซ์แนวทแยงที่มีค่าในแนวทแยง เท่ากับ $p(x_i)(1-p(x_i))$

ผลที่ได้จากการหาอนุพันธ์ของฟังก์ชันล๊อกความเป็นไปได้ จะนำไปใช้ในการประมาณค่าพารามิเตอร์ β ซึ่งมีการนำเสนอวิธีการประมาณค่าอยู่หลายวิธีการ ในที่นี้จะอาศัยวิธีการประมาณค่าที่เรียกว่า วิธีนิวตัน-ราฟสัน (Newton-Raphson Method) ในการประมาณค่า ซึ่งจะใช้ อนุกรมเทเลอร์ (Taylor's series) มาสร้างเป็นสมการหลัก

กำหนดให้

$$l(\beta)=\ln\left[L(\beta)\right] \tag{113}$$

สร้างอนุกรมเทเลอร์ เพื่อประมาณค่า พารามิเตอร์ β ดังนี้

$$l(\beta)\approx l(\beta_{initial})+(\beta-\beta_{initial})l'(\beta_{initial})+\frac{1}{2}(\beta-\beta_{initial})^2 l''(\beta_{initial}) \tag{114}$$

หาค่าสูงสุดของฟังก์ชันล๊อกความเป็นไปได้ ($l(\beta)$ Maximization) โดยการหาอนุพันธ์อันดับหนึ่งจากอนุกรมในสมการ (114) เทียบกับ β จะได้

$$l'(\beta)\approx l'(\beta_{initial})+(\beta-\beta_{initial})l''(\beta_{initial})=0 \tag{115}$$

หรือ

$$\beta=\beta_{initial}-\frac{l'(\beta_{initial})}{l''(\beta_{initial})} \tag{116}$$

นั่นคือ

$$\beta=\beta_{initial}-\left(\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left[L(\beta_{initial})\right]\right)^{-1}\frac{\partial}{\partial\beta}\ln\left[L(\beta_{initial})\right] \tag{117}$$

แทนค่าที่ได้จาก สมการ (110) และสมการ (112) ลงในสมการ (117) จะได้

$$\beta = \beta_{initial} - (X'WX)^{-1} X'(y-p) \qquad (118)$$

ผลจากสมการ (118) จะสามารถประมาณค่าพารามิเตอร์ β ได้จาก ค่า β ณ ตำแหน่งเริ่มต้น (initial) ซึ่งไม่ทราบค่า ดังนั้น เพื่อให้ได้คำตอบที่ใกล้เคียงที่สุด การประมาณค่าพารามิเตอร์ β จึงต้องกำหนดค่าเริ่มต้น ให้กับค่า β ($\beta_{initial}$) แล้วอาศัยวิธีการวนซ้ำ (iteration) แก้สมการ จากสมการ (118) ในการเรียนรู้คำตอบ เพื่อให้ได้ ค่า β ที่ไม่มีการเปลี่ยนแปลงค่า

สรุปวิธีการประมาณค่าพารามิเตอร์ β โดยวิธีนิวตัน-ราฟสันได้ดังนี้

1) กำหนดค่าเริ่มต้นให้กับพารามิเตอร์ β

โดยที่ $\beta_{initial} = \beta_0 = 0$

2) ประมาณค่าพารามิเตอร์ β ณ ตำแหน่งการวนซ้ำที่ t ดังสมการ

$$\beta_t = \beta_{t-1} - (X'W_{t-1}X)^{-1} X'(y-p_{t-1}) \qquad (119)$$

3) ทำการวนซ้ำแก้สมการ (119) จนกระทั่ง ค่าพารามิเตอร์ β ไม่เปลี่ยนแปลง นั่นคือ $\beta_t \approx \beta_{t-1}$ นอกจากนั้นเมตริกซ์ความแปรปรวนร่วมของพารามิเตอร์ β จะเท่ากับ $(X'WX)^{-1}$

ผลที่ได้จากการประมาณค่าพารามิเตอร์ จะช่วยให้เราทราบค่าพารามิเตอร์ β ที่เหมาะสมกับโมเดลการถดถอยโลจิสติก ดังนั้น เมื่อมีเวกเตอร์ของข้อมูลทดสอบ x_i สำหรับใช้ประมาณค่า y_i จะสามารถหาค่าความน่าจะเป็น ที่ y_i มีค่าตามที่กำหนดได้ โดยคำนวณจากฟังก์ชัน โลจิสติก ในสมการ (98) ซึ่งค่าความน่าจะเป็น ($p(x_i)$) ที่ได้จะใช้เป็นตัวตัดสินว่า y_i มีค่าตามที่กำหนดหรือไม่

นั่นคือ ถ้าค่า $p(x_i)$ มากว่าหรือเท่ากับ 0.5 แสดงว่า y_i มีค่าตามที่กำหนด แต่ถ้าไม่ใช่แสดงว่า y_i ไม่ได้มีค่าเป็นไปตามที่กำหนด ดังนั้นเมื่อ y_i เป็นกลุ่มที่กำหนดให้กับข้อมูล ด้วยวิธีการวิเคราะห์ถดถอยโลจิสติก จะสามารถจำแนกกลุ่มของข้อมูลได้ โดยอาศัยค่าความน่าจะเป็นที่ชุดข้อมูล x_i จะอยู่ในกลุ่มที่กำหนดในการพิจารณา

6.2 การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์

แสดงการประยุกต์ การวิเคราะห์การถดถอยโลจิสติก กับข้อมูลดีเอ็นเอไมโครอาร์เรย์ ได้ดังนี้

6.2.1 แหล่งข้อมูลและลักษณะข้อมูล

ข้อมูลที่นำมาใช้ เป็น กรณีศึกษา คือ ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมีย เช่นเดียวกับข้อมูลที่วิเคราะห์ในบทที่ 5 ซึ่งได้อธิบายถึงแหล่งที่มาและลักษณะของข้อมูลแล้วในบทที่ 3 หัวข้อ 3.2.2 และจากตาราง 3.14 แสดงให้เห็นถึงตัวอย่างของข้อมูลที่จะใช้ในการสร้างโมเดลการ

จำแนกประเภท ซึ่งมีทั้งหมด 38 ตัวอย่างข้อมูล โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่ม เอแอลแอล (ALL) 27 ตัวอย่าง และกลุ่มเอเอ็มแอล (AML) 11 ตัวอย่าง สำหรับข้อมูลที่ใช้ในการทดสอบผลการวิเคราะห์จะตัดออกมาจากชุดข้อมูลทั้งหมดซึ่งแสดงไว้ในตาราง 3.13 โดยมีตัวอย่างของข้อมูลที่จะใช้ทดสอบผลการวิเคราะห์จำนวน 34 ตัวอย่าง ซึ่งประเภทของมะเร็งในชุดข้อมูลนี้จะประกอบไปด้วย กลุ่ม เอแอลแอล 20 ตัวอย่าง และกลุ่มเอเอ็มแอล 14 ตัวอย่าง โดยแสดงลักษณะของข้อมูลชุดนี้ในดังตาราง 5.1 ในบทที่ 5

6.2.2 วิธีการวิเคราะห์

1) วิเคราะห์ข้อมูลโดยใช้ฟังก์ชั่นการวิเคราะห์และพัฒนาขึ้นเอง จากโปรแกรม ภาษา R เวอร์ชัน 2.3.1 ในการคำนวณ

2) เตรียมข้อมูลโดยใช้วิธีการเดียวกับการวิเคราะห์ข้อมูลในบทที่ 3 หัวข้อ 3.2.2 ซึ่งผลการวิเคราะห์จะได้ยีนจำนวน 3,051 ยีน ในกลุ่มข้อมูลตัวอย่าง 72 กลุ่มข้อมูล ที่ประกอบไปด้วย ตัวอย่างข้อมูลสำหรับสร้างโมเดล 38 ข้อมูล และสำหรับทดสอบ 34 ข้อมูล เป็นข้อมูลดั้งต้น

3) เลือกตัวแปรสำหรับการจำแนกประเภทข้อมูล โดยวิธีการวิเคราะห์ค่าเอนโทรปี วิธีการวิเคราะห์ค่าความแปรปรวน และวิธีวิเคราะห์องค์ประกอบหลัก สำหรับวิธีการวิเคราะห์องค์ประกอบหลัก จะทำการสร้างข้อมูลชุดใหม่ให้อยู่ในรูปของคะแนนองค์ประกอบหลัก ทั้งข้อมูลที่ใช้ในการสร้างโมเดล และข้อมูลที่จะทดสอบ วิธีการเหล่านี้ได้อธิบายใน บทที่ 5

4) วิเคราะห์การถดถอยโลจิสติก เพื่อสร้างโมเดลการถดถอยโลจิสติก โดยใช้สมการ (98) ทั้งนี้ กำหนดให้ $p(x_i)$ คือความน่าจะเป็นที่ตัวอย่างข้อมูลตัวที่ i จะอยู่ในกลุ่ม เอเอ็มแอล (AML) และ1- $p(x_i)$ คือความน่าจะเป็นที่ตัวอย่างข้อมูลตัวที่ i จะอยู่ในกลุ่ม เอแอลแอล (ALL)นอกจากนี้จะใช้ตัวแปรที่ได้จากขั้นตอนที่ 3 เป็นตัวแปรทำนายค่า ดังนั้นการวิเคราะห์ข้อมูล จึงแบ่งออก เป็น 3 การทดลอง หลักๆ ตามตัวแปรที่ได้ จากกระบวนการเลือกตัวแปรในขั้นตอนที่ 3

5) ทำนาย ค่าความน่าจะเป็น และ กลุ่มของข้อมูล ในตัวอย่างข้อมูลทดสอบ แล้วเปรียบเทียบผลการทำนาย กับ กลุ่มของข้อมูลเดิม

6.2.3 ผลการทดลอง

แสดงผลการทดลองใน 3 การทดลอง ตามวิธีการเลือกตัวแปรสำหรับจำแนกประเภทดังนี้

- **การทดลองที่ 1**

การทดลองนี้ ใช้วิธีการวิเคราะห์ค่าเอนโทรปีในการเลือกตัวแปรที่มีความสำคัญต่อการจำแนกประเภท ซึ่งในที่นี้ก็คือ ยีน แต่จากการวิเคราะห์ และข้อจำกัดของการวิเคราะห์การถดถอยลอจิสติกที่กล่าวไปก่อนหน้านี้ จะใช้ยีนได้ในจำนวนไม่เกิน 37 ยีนในการวิเคราะห์ โดยยีนเหล่านี้จะเป็นยีน ที่มีค่าประมาณของค่าเอนโทรปี E(X) น้อยที่สุด ซึ่งได้แสดงยีนและค่าประมาณของค่าเอนโทรปีเหล่านี้ในตาราง 5.2

ผลจากการวิเคราะห์ถดถอยโลจิสติกในกลุ่มยีนที่เลือก จะได้พารามิเตอร์ β_j แสดงได้ดังตาราง 6.1 ดังนี้

ตาราง 6.1 ค่าพารามิเตอร์ β_j จากการวิเคราะห์การถดถอยโลจิสติก โดยใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด 37 ยีน

Gene Id (X_j)	β_j	Gene Id (X_j)	β_j
Intercept	-35.40	X14008_rna1_f_at	-6.85
M55150_at	-16.56	HG3454-HT3647_at	3.19
U50136_rna1_at	-4.06	Y00787_s_at	-0.34
Y12670_at	6.73	X85116_rna1_s_at	-7.08
U46499_at	4.85	L47738_at	4.09
M77142_at	-14.47	U62136_at	24.19
X95735_at	7.27	L09209_s_at	-5.94
M80254_at	-7.83	X59417_at	-3.62
M23197_at	10.63	M22960_at	-6.30
J04615_at	-4.33	D10495_at	-9.57
U82759_at	13.23	M83652_s_at	2.92
M91432_at	2.35	M63138_at	23.43
M92287_at	-3.32	AFFX-HUMTFRR/M11507_M_at	3.48
M16038_at	9.00	M27891_at	-6.58
M21551_rna1_at	6.66	M83667_rna1_s_at	10.76
U32944_at	-11.05	X17042_at	0.46
U22376_cds2_s_at	-6.50	U67963_at	-12.36
J03801_f_at	13.42	M11147_at	10.35
M19045_f_at	-15.48	U46751_at	-1.25

จากตาราง 6.1 ค่าพารามิเตอร์ที่ได้ จะถูกนำไปใช้ในการหาค่าความน่าจะเป็น ($p(x_i)$) ที่ข้อมูลทดสอบ จะอยู่ในกลุ่มเอเอ็มแอล และความน่าจะเป็นดังกล่าวจะช่วยในการทำนายกลุ่มของข้อมูลเหล่านี้ ซึ่งผลการวิเคราะห์จะแสดงได้ดังตาราง 6.2

ตาราง 6.2 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด จำนวน 37 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	0.00	ALL	56	ALL	0.00	ALL
40	ALL	0.03	ALL	57	AML	0.94	AML
41	ALL	0.00	ALL	59	AML	1.00	AML
42	ALL	0.00	ALL	59	ALL	0.00	ALL
43	ALL	0.00	ALL	*60	AML	0.00	ALL
44	ALL	0.00	ALL	61	AML	1.00	AML
*45	ALL	1.00	AML	62	AML	1.00	AML
46	ALL	0.01	ALL	63	AML	1.00	AML
47	ALL	0.00	ALL	*64	AML	0.00	ALL
48	ALL	0.02	ALL	65	AML	1.00	AML
49	ALL	0.00	ALL	*66	AML	0.00	ALL
50	AML	1.00	AML	*67	ALL	1.00	AML
51	AML	1.00	AML	68	ALL	0.00	ALL
52	AML	1.00	AML	*69	ALL	0.97	AML
*53	AML	0.02	ALL	70	ALL	0.00	ALL
54	AML	1.00	AML	*71	ALL	1.00	AML
55	ALL	0.00	ALL	72	ALL	0.00	ALL
			Correct: 76.47 %				
			Error: 23.53%				

ตาราง 6.2 เป็นผลการวิเคราะห์การถดถอยโลจิสติกในข้อมูลทดสอบซึ่งเป็นผู้ป่วยโรคมะเร็ง จำนวน 34 ตัวอย่าง เพื่อทำนายประเภทของโรคมะเร็งลิวคีเมีย โดยใช้ยีนจำนวน 37 ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุดเป็นตัวแปร

ผลของการทำนายจะแสดงออกมาในลักษณะของความน่าจะเป็นที่ข้อมูลตัวอย่างจะอยู่ในกลุ่มเอเอ็มแอล (Probs of Class='AML') ทั้งนี้ด้วยค่าความน่าจะเป็นที่มากกว่า 0.5 แสดงว่าข้อมูลตัวอย่างอยู่ในกลุ่มเอเอ็มแอล นอกจากนั้นแสดงว่าข้อมูลตัวอย่างอยู่ในกลุ่มเอแอลแอล ซึ่งจากผลการวิเคราะห์ในตารางจะพบว่าความถูกต้องของการทำนาย เมื่อเทียบกับ กลุ่มของข้อมูลที่กำหนดมาแต่เดิมนั้นมีความถูกต้อง 76.47 เปอร์เซ็นต์

จะเห็นว่าผลการทำนายที่ได้ไม่ค่อยดีนักซึ่งสาเหตุนั้นอาจเป็นเพราะว่า จำนวนตัวแปรที่นำมาวิเคราะห์นั้นมีน้อยเกินไป หรืออีกสาเหตุก็เพราะว่ามียีนบางยีนไม่มีคุณสมบัติเป็นตัวทำนายค่าที่ดี แต่ทั้งนี้ด้วยข้อจำกัดของวิธีการ ไม่สามารถใช้จำนวนตัวแปรที่มากกว่านี้ในการวิเคราะห์ได้อีกแล้ว ดังนั้นการพิจารณาเลือกยีนจึงต้องให้ความสำคัญกับการเลือกยีน ที่อยู่ภายในข้อจำกัดเหล่านี้แต่เป็นยีนที่มีคุณสมบัติเป็นตัวทำนายค่าที่ดี การวิเคราะห์ในขั้นต่อไปจึงทำการพิจารณายีน เป็นช่วงๆ แล้วทำการ

วิเคราะห์กับยีนเหล่านี้ เพื่อวัดผลการทำนาย และแสดงผลสรุปของการทำนายกลุ่มของข้อมูลได้ดังตาราง 6.3

ตาราง 6.3 ผลสรุปของการวิเคราะห์การถดถอยโลจิสติกของกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด 37 ช่วงยีน

ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
1-1	26.47	73.53
1-2	23.53	76.47
1-3	32.35	67.65
1-4	32.35	67.65
1-5	29.41	70.59
1-6	26.47	73.53
1-7	32.35	67.65
1-8	20.59	79.41
1-9	11.76	88.24
1-10	11.76	88.24
1-11	14.71	85.29
1-12	8.82	91.18
1-13	11.76	88.24
1-14	11.76	88.24
1-15	14.71	85.29
1-16	17.65	82.35
1-17	14.71	85.29
1-18	8.82	91.18
1-19	14.71	85.29
1-20	17.65	82.35
1-21	20.59	79.41
1-22	17.65	82.35
1-23	20.59	79.41
1-24	20.59	79.41
1-25	20.59	79.41
1-26	20.59	79.41
1-27	20.59	79.41
1-28	20.59	79.41
1-29	20.59	79.41
1-30	5.88	94.12
1-31	11.76	88.24
1-32	11.76	88.24
1-33	11.76	88.24
1-34	11.76	88.24
1-35	17.65	82.35
1-36	23.53	76.47
1-37	23.53	76.47

จากตาราง 6.3 พบว่าผลการทำนายที่ได้มีเปอร์เซ็นต์การทำนายกลุ่มข้อมูลถูกอยู่ในระดับที่สูง แต่ไม่มีความแน่นอนโดยในแต่ละช่วงของยีนนั้น จะให้ผลการทำนายที่แตกต่างกัน ซึ่งไม่สามารถบอกได้ว่าช่วงไหนดีที่สุด แม้ว่าเมื่อใช้ยีน 30 ตัวแรกที่มีค่าประมาณของเอนโทรปีน้อยที่สุดในการวิเคราะห์ถดถอยโลจิสติก โดยผลการทำนายกลุ่มข้อมูลทดสอบมีความถูกต้องถึง 94.12 เปอร์เซ็นต์ซึ่งถือว่ามากที่สุด ทั้ง 35 ช่วงยีน แต่ผลที่ได้ยังไม่ได้เปรียบเทียบกับช่วงยีนอื่นๆ นอกจากที่นำเสนอในตารางเช่น ช่วงยีน 2-30 เป็นต้น นอกจากนี้ด้วยจำนวนของยีนที่แตกต่างกันในแต่ละช่วงก็อาจจะมีผลต่อการทำนายค่าด้วยเหมือนกัน ด้วยเหตุนี้จึงสรุปไม่ได้ยีนที่มีค่าประมาณเอนโปรปีต่ำที่สุดจะใช้เป็นตัวแปรจำแนกกลุ่มข้อมูลได้ดีที่สุด

จากยีนในช่วง 1-30 ยีนเมื่อวิเคราะห์ถดถอยโลจิสติก จะแสดงผลจากการจำแนกกลุ่มข้อมูลเหล่านี้ดังตาราง 6.4

ตาราง 6.4 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด จำนวน 30 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	0.00	ALL	56	ALL	0.00	ALL
40	ALL	0.01	ALL	57	AML	0.52	AML
41	ALL	0.00	ALL	59	AML	0.96	AML
42	ALL	0.00	ALL	59	ALL	0.00	ALL
43	ALL	0.00	ALL	*60	AML	0.01	ALL
44	ALL	0.00	ALL	61	AML	1.00	AML
45	ALL	0.00	ALL	62	AML	1.00	AML
46	ALL	0.00	ALL	63	AML	1.00	AML
47	ALL	0.00	ALL	64	AML	0.96	AML
48	ALL	0.00	ALL	65	AML	0.62	AML
49	ALL	0.00	ALL	66	AML	0.74	AML
50	AML	1.00	AML	*67	ALL	1.00	AML
51	AML	1.00	AML	68	ALL	0.00	ALL
52	AML	1.00	AML	69	ALL	0.33	ALL
53	AML	0.73	AML	70	ALL	0.00	ALL
54	AML	1.00	AML	71	ALL	0.26	ALL
55	ALL	0.00	ALL	72	ALL	0.00	ALL
			Correct: 94.12 %				
			Error: 5.88%				

จากตาราง 6.4 พบว่าทำนายกลุ่มข้อมูล ผิดพลาดเพียง 2 ข้อมูลตัวอย่างเท่านั้นเองนั่นคือ ตัวอย่างข้อมูลผู้ป่วยคนที่ 60 และ 63 ซึ่งคิดเป็นเปอร์เซ็นต์ผิดพลาดเพียง 5.88 เปอร์เซ็นต์

เพื่อที่จะเปรียบเทียบผลกับกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีมากที่สุด การวิเคราะห์จึงเลือกยีนที่ให้ค่าประมาณของค่าเอนโทรปี มากที่สุดจำนวน 36 ยีน มาเป็นตัวแปรทำนายค่า ซึ่งแสดงผลการวิเคราะห์ได้ดังตาราง 6.5

ตาราง 6.5 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีมากที่สุด จำนวน 36 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
*39	ALL	1.00	AML	*56	ALL	1.00	AML
40	ALL	0.00	ALL	57	AML	1.00	AML
*41	ALL	1.00	AML	59	AML	1.00	AML
42	ALL	0.00	ALL	59	ALL	0.00	ALL
*43	ALL	1.00	AML	60	AML	1.00	AML
*44	ALL	1.00	AML	*61	AML	0.00	ALL
45	ALL	0.00	ALL	*62	AML	0.00	ALL
*46	ALL	1.00	AML	*63	AML	0.00	ALL
*47	ALL	1.00	AML	*64	AML	0.00	ALL
*48	ALL	1.00	AML	*65	AML	0.00	ALL
*49	ALL	1.00	AML	66	AML	1.00	AML
*50	AML	0.00	ALL	67	ALL	0.00	ALL
*51	AML	0.00	ALL	68	ALL	0.00	ALL
*52	AML	0.00	ALL	69	ALL	0.00	ALL
*53	AML	0.00	ALL	*70	ALL	1.00	AML
54	AML	1.00	AML	*71	ALL	1.00	AML
*55	ALL	1.00	AML	72	ALL	0.00	ALL
				Correct: 38.24 %			
				Error: 61.76 %			

จากตาราง 6.5 ผลการทำนายที่ได้ผิดพลาดถึง 61.76 เปอร์เซ็นต์ นั่นแสดงให้เห็นอย่างเด่นชัดว่า กลุ่มยีนที่เลือกมาไม่เหมาะสมกับการใช้เป็นตัวแปรทำนายค่าในการวิเคราะห์ถดถอยโลจิสติก

ผลที่ได้ทั้งหมดจึงสรุปได้ว่า กลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีต่ำ จะเป็นยีนที่ใช้เป็นตัวแปรทำนายค่า ในการวิเคราะห์การถดถอยโลจิสติก ได้เหมาะสมกว่า กลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีที่สูง แต่จำนวนของยีนที่เหมาะสมนั้น จะต้องมีการศึกษาวิเคราะห์ต่อไป

- **การทดลองที่ 2**

การทดลองนี้ ใช้วิธีการวิเคราะห์ค่าความแปรปรวนในการเลือกตัวแปรที่มีความสำคัญต่อการจำแนกประเภท ซึ่งในที่นี้ก็คือ ยีน แต่จากการวิเคราะห์ และข้อจำกัดของการวิเคราะห์การถดถอยลอจิสติกที่ กล่าวไปก่อนหน้านี้ จะใช้ยีนได้ในจำนวนไม่เกิน 37 ยีนในการวิเคราะห์ โดยยีนเหล่านี้จะเป็นยีน ที่มีค่าสัดส่วนของค่าความแปรปรวน V(X) มากที่สุด ซึ่งได้แสดงยีนและค่าสัดส่วนของค่าความแปรปรวน เหล่านี้ ในตาราง 5.6

ผลการวิเคราะห์การถดถอยโลจิสติกโดยใช้ยีนทั้ง 37 ตัวนี้เป็นตัวแปรทำนายค่าจะได้ผลการทำนายกลุ่มของตัวอย่างข้อมูลดังตารางต่อไปนี้

ตาราง 6.6 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุด จำนวน 37 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	0.00	ALL	56	ALL	0.00	ALL
*40	ALL	1.00	AML	*57	AML	0.00	ALL
41	ALL	0.00	ALL	*59	AML	0.00	ALL
42	ALL	0.00	ALL	59	ALL	0.00	ALL
*43	ALL	1.00	AML	*60	AML	0.00	ALL
44	ALL	0.00	ALL	*61	AML	0.00	ALL
45	ALL	0.00	ALL	*62	AML	0.00	ALL
*46	ALL	1.00	AML	*63	AML	0.00	ALL
*47	ALL	1.00	AML	*64	AML	0.00	ALL
*48	ALL	1.00	AML	*65	AML	0.00	ALL
49	ALL	0.00	ALL	*66	AML	0.00	ALL
*50	AML	0.00	ALL	67	ALL	0.00	ALL
*51	AML	0.00	ALL	*68	ALL	1.00	AML
*52	AML	0.00	ALL	69	ALL	0.01	ALL
53	AML	1.00	AML	70	ALL	0.00	ALL
54	AML	1.00	AML	71	ALL	0.00	ALL
*55	ALL	1.00	AML	*72	ALL	1.00	AML
				Correct: 41.18%			
				Error: 58.82%			

ผลจากตาราง 6.6 ทำนายกลุ่มข้อมูลได้ถูกต้องเพียง 41.18 เปอร์เซ็นต์ ถือว่าให้ผลการทำนายไม่ดีนัก จึงต้องทำการพิจารณาที่จำนวนยีนซึ่งน้อยกว่า 37 ยีน โดยแบ่งเป็นช่วงยีนในการวิเคราะห์ และจะแสดงสรุปผลการวิเคราะห์ช่วงยีนเหล่านี้ได้ดังตาราง 6.7

ตาราง 6.7 ผลสรุปของการวิเคราะห์การถดถอยโลจิสติกของกลุ่มยีนที่ค่าสัดส่วนของค่าความแปรปรวนมากที่สุด 37 ช่วงยีน

ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
1-1	5.88	94.12
1-2	5.88	94.12
1-3	8.82	91.18
1-4	11.76	88.24
1-5	11.76	88.24
1-6	11.76	88.24
1-7	8.82	91.18
1-8	0.00	100.00
1-9	5.88	94.12
1-10	5.88	94.12
1-11	5.88	94.12
1-12	8.82	91.18
1-13	2.94	97.06
1-14	11.76	88.24
1-15	5.88	94.12
1-16	5.88	94.12
1-17	5.88	94.12
1-18	5.88	94.12
1-19	2.94	97.06
1-20	5.88	94.12
1-21	5.88	94.12
1-22	2.94	97.06
1-23	5.88	94.12
1-24	5.88	94.12
1-25	20.59	79.41
1-26	20.59	79.41
1-27	5.88	94.12
1-28	5.88	94.12
1-29	17.65	82.35
1-30	29.41	70.59
1-31	23.53	76.47
1-32	32.35	67.65
1-33	38.24	61.76
1-34	47.06	52.94
1-35	47.06	52.94
1-36	17.65	82.35
1-37	58.82	41.18

จากตาราง 6.7 พบว่าผลการทำนายที่ได้มีเปอร์เซ็นต์การทำนายกลุ่มข้อมูลถูก อยู่ในระดับที่สูง โดยเฉพาะช่วงยีนแรกๆ แต่ไม่มีความแน่นอนโดยในแต่ละช่วงของยีนนั้น จะให้ผลการทำนายที่

แตกต่างกัน ซึ่งไม่สามารถบอกได้ว่าช่วงไหนดีที่สุด แม้ว่าเมื่อใช้ยีน 8 ตัวแรกที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุดในการวิเคราะห์ โดยผลการทำนายกลุ่มข้อมูลทดสอบ มีความถูกต้องถึง 100 เปอร์เซ็นต์ซึ่งถือว่าดีที่สุด แต่ผลที่ได้ยังไม่ได้เปรียบเทียบกับช่วงยีนอื่นๆ นอกจากที่นำเสนอในตาราง นอกจากนี้ด้วยจำนวนของยีนที่แตกต่างกันในแต่ละช่วงก็อาจจะมีผลต่อการทำนายค่าด้วยเหมือนกัน การเปรียบเทียบด้วยจำนวนยีนที่ต่างกันจึงไม่สมเหตุสมผล ด้วยเหตุนี้จึงสรุปไม่ได้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุดจะใช้เป็นตัวแปรจำแนกกลุ่มข้อมูลได้ดีที่สุด

จากยีนในช่วง 1-8 ยีนเมื่อวิเคราะห์ถดถอยโลจิสติก จะแสดงผลจากการจำแนกกลุ่มข้อมูลเหล่านี้ดังตาราง 6.8

ตาราง 6.8 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุด จำนวน 8 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	0.00	ALL	56	ALL	0.00	ALL
40	ALL	0.00	ALL	57	AML	1.00	AML
41	ALL	0.00	ALL	59	AML	1.00	AML
42	ALL	0.32	ALL	59	ALL	0.00	ALL
43	ALL	0.00	ALL	60	AML	0.64	AML
44	ALL	0.00	ALL	61	AML	0.99	AML
45	ALL	0.00	ALL	62	AML	1.00	AML
46	ALL	0.00	ALL	63	AML	1.00	AML
47	ALL	0.00	ALL	64	AML	1.00	AML
48	ALL	0.00	ALL	65	AML	0.96	AML
49	ALL	0.00	ALL	66	AML	0.64	AML
50	AML	1.00	AML	67	ALL	0.39	ALL
51	AML	1.00	AML	68	ALL	0.00	ALL
52	AML	1.00	AML	69	ALL	0.00	ALL
53	AML	1.00	AML	70	ALL	0.00	ALL
54	AML	1.00	AML	71	ALL	0.00	ALL
55	ALL	0.00	ALL	72	ALL	0.00	ALL
		Correct: 100 %					
		Error: 0%					

จากตาราง 6.8 จากผลการทำนายกลุ่มข้อมูล พบว่าถูกต้องถึง 100 เปอร์เซ็นต์

เพื่อที่จะเปรียบเทียบผลกับกลุ่มยีนที่มีค่าสัดส่วนของค่าความแปรปรวนน้อยที่สุด การวิเคราะห์จึงเลือกยีนที่มีค่าสัดส่วนของค่าความแปรปรวนน้อยที่สุดจำนวน 36 ยีน มาเป็นตัวแปรทำนายค่า ซึ่งแสดงผลการวิเคราะห์ได้ดังตาราง 6.9

ตาราง 6.9 ผลของการวิเคราะห์ถดถอยโลจิสติกโดยใช้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนน้อยที่สุด จำนวน 36 ยีน เป็นตัวแปรทำนายค่า

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	1.00	AML	56	ALL	0.00	ALL
40	ALL	1.00	AML	57	AML	0.88	AML
41	ALL	1.00	AML	59	AML	0.00	ALL
42	ALL	0.00	ALL	59	ALL	1.00	AML
43	ALL	0.00	ALL	60	AML	0.47	ALL
44	ALL	1.00	AML	61	AML	1.00	AML
45	ALL	0.00	ALL	62	AML	0.00	ALL
46	ALL	1.00	AML	63	AML	1.00	AML
47	ALL	1.00	AML	64	AML	0.00	ALL
48	ALL	1.00	AML	65	AML	0.00	ALL
49	ALL	0.00	ALL	66	AML	1.00	AML
50	AML	0.00	ALL	67	ALL	0.00	ALL
51	AML	1.00	AML	68	ALL	1.00	AML
52	AML	1.00	AML	69	ALL	1.00	AML
53	AML	1.00	AML	70	ALL	1.00	AML
54	AML	1.00	AML	71	ALL	0.98	AML
55	ALL	1.00	AML	72	ALL	1.00	AML
			Correct: 41.18%				
			Error: 58.82 %				

จากตาราง 6.9 ผลการทำนายที่ได้ผิดพลาดถึง 58.82 เปอร์เซ็นต์ นั่นแสดงให้เห็นอย่างเด่นชัดว่า กลุ่มยีนที่เลือกมาไม่เหมาะสมกับการใช้เป็นตัวแปรทำนายค่าในการวิเคราะห์ถดถอยโลจิสติก

ผลที่ได้ทั้งหมดจึงสรุปได้ว่า กลุ่มยีนที่มีสัดส่วนของค่าความแปรปรวนมากจะเป็นยีนที่ใช้เป็นตัวแปรทำนายค่า ในการวิเคราะห์การถดถอยโลจิสติก ได้เหมาะสมกว่า กลุ่มยีนที่มีค่าสัดส่วนของค่าความแปรปรวนน้อย แต่สรุปไม่ได้ว่ายีนที่มีค่าสัดส่วนของค่าความแปรปรวนสูงที่สุดจะใช้เป็นตัวแปรจำแนกกลุ่มข้อมูลได้ดีที่สุด เนื่องจากจะต้องมีการพิจารณาจำนวนของยีนที่เหมาะสม และ การเลือกยีนที่มีความสำคัญต่อการจำแนกกลุ่มข้อมูลได้ดีที่สุดในอีกขั้นตอนหนึ่ง จึงจะให้ผลการวิเคราะห์ที่ดีได้ ซึ่งในส่วนนี้จำเป็นที่จะต้องมีการศึกษาต่อไป

- **การทดลองที่ 3**

โดยวิธีนี้ ตัวแปรที่ใช้สำหรับการจำแนกประเภทจะต่างกับ 2 วิธีที่ผ่านมา ซึ่งใช้ยีนเป็นตัวแปร แต่ในการทดลองนี้ ตัวแปรที่ใช้ในการวิเคราะห์ คือ องค์ประกอบหลักของข้อมูล ทั้งนี้องค์ประกอบหลักของข้อมูล ถือว่าเป็นตัวแปรที่สร้างขึ้นใหม่แทนตัวแปร ยีน อย่างเดิม และการสร้างองค์ประกอบหลัก จะอาศัยวิธีการวิเคราะห์องค์ประกอบหลัก ในการหาองค์ประกอบหลักของยีนทั้ง 3,051 ยีน นอกจากนี้ ชุดข้อมูลที่จะใช้ในการสร้างองค์ประกอบหลัก จะเป็นชุดข้อมูลที่ใช้ในการสร้างโมเดลการวิเคราะห์การจำแนกประเภท ซึ่งมีจำนวน 38 ตัวอย่าง

ผลจากการวิเคราะห์องค์ประกอบหลัก ทำให้ได้องค์ประกอบหลักที่เป็นตัวแทนของยีนทั้งหมดด้วยค่าความแปรปรวนต่างๆ แสดงอยู่ในตาราง 5.9 ซึ่งผลของการวิเคราะห์องค์จะกอบหลัก จะได้ค่าคะแนนองค์ประกอบหลักจำนวน 36 องค์ประกอบหลักแสดงในตาราง 5.10 และตาราง 5.11 จากค่าคะแนนองค์ประกอบหลักเหล่านี้ เมื่อนำมาใช้ในการวิเคราะห์การถดถอยโลจิสติก จะสามารถทำนายกลุ่มของตัวอย่างข้อมูลทดสอบใน 34 ตัวอย่าง โดยใช้ 36 องค์ประกอบหลักเป็นตัวแปร และแสดงผลการวิเคราะห์ได้ดังตาราง 6.10

ตาราง 6.10 ผลการจำแนกประเภทข้อมูล จากการวิเคราะห์การถดถอยโลจิสติกโดยใช้ 36 องค์ประกอบหลักแรก เป็นตัวแปร

Samples	Original Class	Results of LRA		Samples	Original Class	Results of LRA	
		Probs Of Class='AML'	Class Estimation			Probs Of Class='AML'	Class Estimation
39	ALL	0.00	ALL	56	ALL	0.00	ALL
40	ALL	0.00	ALL	57	AML	1.00	AML
41	ALL	0.00	ALL	59	AML	1.00	AML
42	ALL	0.01	ALL	59	ALL	0.00	ALL
43	ALL	0.00	ALL	60	AML	0.99	AML
44	ALL	0.00	ALL	61	AML	0.99	AML
45	ALL	0.00	ALL	62	AML	1.00	AML
46	ALL	0.00	ALL	63	AML	1.00	AML
47	ALL	0.00	ALL	64	AML	1.00	AML
48	ALL	0.00	ALL	65	AML	1.00	AML
49	ALL	0.00	ALL	*66	AML	0.27	ALL
50	AML	1.00	AML	*67	ALL	0.51	AML
51	AML	1.00	AML	68	ALL	0.00	ALL
52	AML	1.00	AML	69	ALL	0.00	ALL
53	AML	1.00	AML	70	ALL	0.00	ALL
54	AML	1.00	AML	71	ALL	0.00	ALL
55	ALL	0.00	ALL	72	ALL	0.00	ALL
			Correct: 94.12 %				
			Error: 5.88 %				

จากตาราง 6.10 แสดงให้เห็นผลของการทำนายกลุ่มตัวอย่างข้อมูล พบว่า กลุ่มข้อมูลที่ทำนายนั้นทำนายผิดจากกลุ่มที่กำหนดมาตอนต้นเพียง 2 ตัวอย่างข้อมูล หรือคิดเป็น 5.88 เปอร์เซ็นต์ และเมื่อพิจารณาทุกๆช่วงขององค์ประกอบหลักใน 36 องค์ประกอบจะแสดงผลการวิเคราะห์ดังตาราง 6.11

ตาราง 6.11 ผลสรุปของการวิเคราะห์การถดถอยโลจิสติกในชุดข้อมูลโดยอาศัยองค์ประกอบหลักเป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 36 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
1-1	16.19	41.18	58.82
1-2	29.89	11.76	88.24
1-3	37.77	8.82	91.18
1-4	43.74	8.82	91.18
1-5	48.49	11.76	88.24
1-6	52.33	8.82	91.18
1-7	55.67	5.88	94.12
1-8	58.85	8.82	91.18
1-9	61.54	2.94	97.06
1-10	64.13	5.88	94.12
1-11	66.41	5.88	94.12
1-12	68.56	5.88	94.12
1-13	70.47	5.88	94.12
1-14	72.33	5.88	94.12
1-15	74.13	2.94	97.06
1-16	75.79	2.94	97.06
1-17	77.42	2.94	97.06
1-18	78.97	2.94	97.06
1-19	80.41	2.94	97.06
1-20	81.80	2.94	97.06
1-21	83.17	5.88	94.12
1-22	84.50	5.88	94.12
1-23	85.78	5.88	94.12
1-24	87.01	2.94	97.06
1-25	88.22	5.88	94.12
1-26	89.39	5.88	94.12
1-27	90.53	2.94	97.06
1-28	91.64	2.94	97.06
1-29	92.72	2.94	97.06
1-30	93.78	2.94	97.06
1-31	94.80	2.94	97.06
1-32	95.73	2.94	97.06
1-33	96.65	2.94	97.06
1-34	97.53	5.88	94.12
1-35	98.39	2.94	97.06
1-36	99.23	5.88	94.12

จากตาราง 6.11 ผลการทำนายการวิเคราะห์การจำแนกประเภท โดยใช้ช่วงขององค์ประกอบหลักแรกๆ เป็นตัวแปร พบว่าให้ผลการทำนายที่มีเปอร์เซ็นต์การทำนายถูกตามกลุ่มที่กำหนดมาตอนต้นในระดับที่สูง แม้ว่าที่ 1 หรือ 2 องค์ประกอบหลักแรกจะให้ผลการทำนายที่มีเปอร์เซ็นต์การทำนายผิดที่สูงอยู่ก็ตาม ซึ่งสาเหตุอาจจะเป็นเพราะว่าความแปรปรวนของข้อมูลในองค์ประกอบหลักทั้ง 2 องค์ประกอบหลักนั้นมีค่าน้อย ซึ่งมีผลทำให้องค์ประกอบหลักทั้ง 2 เป็นตัวแทนของข้อมูลทั้งหมดไม่ดีนัก การใช้องค์ประกอบหลักเพียงแค่ 1 หรือ 2 องค์ประกอบนี้ในการวิเคราะห์จึงให้ผลการทำนายที่ผิดพลาดมากดังกล่าว

นอกจากนี้ ผลจากการทำนายกลุ่มของตัวอย่างข้อมูล โดยใช้ช่วงขององค์ประกอบหลัก ที่มีค่าความแปรปรวนของข้อมูลน้อยๆ เป็นตัวแปร จะแสดงผลการวิเคราะห์ได้ดังตาราง 6.12

ตาราง 6.12 ผลสรุปของการวิเคราะห์การถดถอยโลจิสติกในชุดข้อมูลโดยอาศัยองค์ประกอบหลักที่มีค่าความแปรปรวนน้อยๆ เป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 20 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
36-36	0.84	41.18	58.82
35-36	1.70	41.18	58.82
34-36	2.58	41.18	58.82
33-36	3.50	41.18	58.82
32-36	4.43	41.18	58.82
31-36	5.44	41.18	58.82
30-36	6.51	41.18	58.82
29-36	7.59	41.18	58.82
28-36	8.69	41.18	58.82
27-36	9.83	41.18	58.82
26-36	11.00	41.18	58.82
25-36	12.21	41.18	58.82
24-36	13.44	41.18	58.82
23-36	14.72	41.18	58.82
22-36	16.05	41.18	58.82
21-36	17.42	41.18	58.82
20-36	18.81	41.18	58.82
19-36	20.25	41.18	58.82
18-36	21.80	41.18	58.82
17-36	23.43	41.18	58.82

จากตาราง 6.12 แสดงให้เห็นผลการทำนายกลุ่มข้อมูล ที่มีเปอร์เซ็นต์การทำนายที่ผิดจากข้อมูลเดิมอยู่ในระดับที่สูง

ข้อสรุปจากตาราง 6.11 และ 6.12 สรุปได้ว่า องค์ประกอบหลัก ในระดับแรกๆ ซึ่งมีค่าความแปรปรวนของข้อมูลสูงมากสามารถใช้เป็นตัวแปรทำนายในการวิเคราะห์การถดถอยโลจิสติกได้ดีกว่าองค์ประกอบหลักที่มีความแปรปรวนของข้อมูลต่ำ อย่างไรก็ตาม จากตาราง 6.11 จะพบว่ามีบางช่วงองค์ประกอบ ที่แม้จะให้ค่าความแปรปรวนของข้อมูลสูง แต่เปอร์เซ็นต์การทำนายผิด ก็ยังมากกว่าช่วงองค์ประกอบหลักอื่นที่มีค่าความแปรปรวนสะสมน้อยกว่า การหาว่า องค์ประกอบไหน หรือจำนวนองค์ประกอบจำนวนเท่าใด เหมาะสมเป็นตัวแปรทำนายการจำแนกประเภทนั้น เป็นสิ่งที่ต้องศึกษาต่อไป

6.3 วิจารณ์และสรุปผล

แนวทางการประยุกต์วิธีวิเคราะห์การถดถอยแบบโลจิสติกกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ในงานวิจัยนี้คือ ใช้การวิเคราะห์การถดถอยแบบโลจิสติก วิเคราะห์ข้อมูลดีเอ็นเอไมโครเรย์ของมะเร็งลิวคีเมีย เพื่อทำนายประเภทโรคมะเร็งของตัวอย่างข้อมูลในชุดข้อมูลทดสอบ โดยใช้ยีนเป็นตัวแปรเช่นเดียวกับการวิเคราะห์การจำแนกประเภท ข้อดีของวิธีการนี้เมื่อเทียบกับวิธีการวิเคราะห์การจำแนกประเภทคือ ช่วยอธิบายความหมายของความแตกต่างของกลุ่มข้อมูล ได้ง่ายกว่าการวิเคราะห์การจำแนกประเภทข้อมูลเนื่องจากสามารถให้ผลการทำนายเป็นค่าความน่าจะเป็นที่ตัวอย่างข้อมูลจะเป็นมะเร็งชนิดใดๆ นอกจากนี้ ในกรณีข้อมูลที่นำมาใช้วิเคราะห์เป็นค่าไม่ต่อเนื่องก็สามารถนำมาวิเคราะห์ได้ แต่ข้อเสียของวิธีนี้ที่เห็นได้อย่างชัดเจนกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งลิวคีเมีย คือไม่สามารถที่จะวิเคราะห์กับตัวแปรที่มีจำนวนมากกว่าตัวอย่างข้อมูลได้ จำนวนตัวแปรจึงจำกัดอยู่ที่จำนวนไม่เกินจำนวนของกลุ่มตัวอย่าง ดังนั้นวิธีการคัดเลือกตัวแปรที่สำคัญต่อการทำนายจึงนำมาใช้วิเคราะห์ข้อมูลก่อนที่จะวิเคราะห์การถดถอยโลจิสติก วิธีการดังกล่าวได้แก่ วิธีการวิเคราะห์ค่าเอนโทรปี วิธีวิเคราะห์ค่าความแปรปรวน และวิธีวิเคราะห์องค์ประกอบหลัก ผลการวิเคราะห์ที่ได้นำไปเปรียบเทียบผลการทำนายกับประเภทของตัวอย่างข้อมูลในชุดข้อมูลทดสอบที่กำหนดมาตอนต้น เช่นเดียวกับการวิเคราะห์การจำแนกประเภท

ผลการวิเคราะห์ที่ได้ทั้ง 3 ลักษณะ สรุปได้ดังนี้

1) ผลการวิเคราะห์การถดถอยแบบโลจิสติก โดยใช้วิธีการวิเคราะห์ค่าเอนโทรปีในการเลือกยีน เป็นตัวแปรทำนายกลุ่มข้อมูล พบว่าการวิเคราะห์การถดถอยแบบโลจิสติก จะให้ผลการทำนายกลุ่มข้อมูลได้ดี เมื่อยีนที่นำมาใช้เป็นตัวแปรทำนายค่าเป็นยีนที่มีค่าประมาณของเอนโทรปีต่ำที่สุดด้วยจำนวนยีนที่เหมาะสม ซึ่งผลการวิเคราะห์ เมื่อใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุดจำนวน 30 ยีน เป็นตัวแปรจำแนกประเภท พบว่าผลการทำนายกลุ่มข้อมูลมีความถูกต้องถึง 99.12 เปอร์เซ็นต์

แต่ด้วยข้อจำกัดของวิธีการการถดถอยโลจิสติก ทำให้ยีนที่นำมาใช้เป็นตัวแปรทำนายกลุ่มข้อมูลนั้น มีจำนวนน้อย ผลการวิเคราะห์ที่ได้จึงไม่มีความน่าเชื่อถือ

2) ผลการวิเคราะห์การถดถอยแบบโลจิสติก โดยใช้วิธีการวิเคราะห์ค่าความแปรปรวนในการเลือกยีน สำหรับเป็นตัวแปรทำนายกลุ่มข้อมูล พบว่า การวิเคราะห์การถดถอยแบบโลจิสติกจะให้ผลการทำนายกลุ่มข้อมูลได้ดีเมื่อยีนที่นำมาใช้เป็นตัวแปรทำนายค่า เป็นยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุดด้วยจำนวนยีนที่เหมาะสม ซึ่งผลการวิเคราะห์ที่ได้ เมื่อใช้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุดจำนวน 8 ยีน พบว่าผลการทำนายกลุ่มข้อมูลมีความถูกต้องทั้งหมด 100 เปอร์เซ็นต์ แต่ข้อจำกัดของวิธีการนี้ก็คือ ไม่สามารถระบุจำนวนของยีนที่เหมาะสมได้ นอกจากนี้ ด้วยข้อจำกัดของวิธีการการถดถอยโลจิสติก ทำให้ยีนที่นำมาใช้เป็นตัวแปรทำนายกลุ่มข้อมูลนั้น มีจำนวนน้อย ผลการวิเคราะห์ที่ได้จึงไม่มีความน่าเชื่อถือ

3) ผลการวิเคราะห์การถดถอยแบบโลจิสติก โดยอาศัยตัวแปรทำนายค่า ที่สร้างขึ้นด้วยวิธีการวิเคราะห์องค์ประกอบหลัก พบว่า องค์ประกอบหลักที่มีค่าความแปรปรวนมากๆ จะสามารถใช้เป็นตัวแปรทำนายค่าได้ ซึ่งผลการวิเคราะห์โดยใช้องค์ประกอบหลักดังกล่าว จะให้ผลการวิเคราะห์ที่มีความถูกต้องสูง ซึ่งจากการวิเคราะห์ที่ 36 องค์ประกอบหลักซึ่งมีความแปรปรวน 99.23 เปอร์เซ็นต์จะให้ผลการทำนายกลุ่มข้อมูลที่ถูกต้องถึง 94.12 เปอร์เซ็นต์ ทั้งนี้ข้อดีของการวิเคราะห์องค์ประกอบหลักก็คือตัวแปรที่ใช้ในการทำนายกลุ่มของข้อมูลนั้น ถือได้ว่าเป็นตัวแทนของตัวแปรในข้อมูลทั้งหมด ด้วยค่าความแปรปรวนที่สูง ปัญหาในเรื่องข้อจำกัดของจำนวนยีน หรือจำนวนตัวแปรที่เหมาะสม นั้นจึงไม่มี

จากผลการวิเคราะห์จำแนกประเภททั้ง 3 การทดลองจะพบว่า เทคนิคการการถดถอยแบบโลจิสติก จะให้ผลการวิเคราะห์ที่ดีหรือไม่นั้นอยู่ที่ ตัวแปรทำนายค่า ซึ่งถ้าก่อนการวิเคราะห์มีการเลือกตัวแปรทำนายค่าที่ดีแล้ว ผลการทำนายกลุ่มของข้อมูลย่อมให้ผลการทำนายที่ดีตามไปด้วย ทั้งนี้ นอกจากเทคนิควิธีการที่ใช้ในการเลือกตัวแปรจำแนกประเภทดังที่นำเสนอยังมีหลายวิธีการที่นำเสนอในผลงานวิจัยต่างๆ ที่จำเป็นจะต้องมีการวิเคราะห์เปรียบเทียบผล เช่นเดียวกับการวิเคราะห์การจำแนกประเภท