การเปรียบเทียบวิธีการของการวิเคราะห์ความสำคัญของกลุ่มยีนและการถดถอยลอจิสติกทวิภาคสำหรับการตรวจสอบความสัมพันธ์ระหว่างกลุ่มยีนและฟีโนไทป์ทวิภาค

A Method Comparison of Gene Set Enrichment Analysis and Binary Logistic Regression for Investigating the Relationship between Gene Sets and a Binary Phenotype

สุธิภาส สิงห์เรือง* และวิฐรา พึ่งพาพงศ์

ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

แขวงวังใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330

Sutipat Singruang* and Vitara Pungpapong

Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University,

Wangmai, Pathumwan, Bangkok 10330

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและการถดถอยลอจิสติกทวิภาค ในการหาค่าพี (p-value) ของแต่ละกลุ่มยีน โดยคำนึงถึงความสัมพันธ์และการทำงานร่วมกันเป็นกลุ่มของยีน โดยการศึกษานี้จะเปรียบเทียบประสิทธิภาพ จากการวิเคราะห์ข้อมูลจำลองทั้งในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนของยีนหรือตัวแปรอิสระ และกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ หรือที่เรียกว่าข้อมูลที่มีมิติสูง ในขอบเขตการศึกษาต่าง ๆ กัน ในงานวิจัยนี้จะเปรียบเทียบค่าอัตราความผิดพลาดรวม และกำลังการทดสอบเพื่อวัดประสิทธิภาพจากวิธีทั้งสอง จากการศึกษาภายใต้ขอบเขตดังกล่าว ผลปรากฏว่าวิธีการถดถอยลอจิสติกทวิภาค มีกำลังการทดสอบ (เฉลี่ย) สูงในกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ ในขณะที่วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีกำลังการทดสอบ (เฉลี่ย) สูงในกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ แต่เมื่อพิจารณาถึงการวัดประสิทธิภาพจากค่าอัตราความผิดพลาดรวม พบว่าวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าต่ำสำหรับกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ ในขณะที่วิธีการถดถอยลอจิสติกทวิภาคมีค่าต่ำสำหรับกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ

คำสำคัญ : การวิเคราะห์ความสำคัญของกลุ่มยีน; การถดถอยลอจิสติกทวิภาค; วิธีแลสโซ่; อัตราความผิดพลาดรวม; กำลังการทดสอบ

***ผู้รับผิดชอบบทความ :** sutipat.sr@gmail.com

Abstract

This research is aimed to study and compare gene set enrichment analysis method and binary logistic regression in finding p-values of each gene set. Here we consider the relationship and collaboration among genes in each gene set. In this study, the performance of two methods are compared using simulated data in two cases: (i) sample size is larger than the number of genes or independent variables (ii) sample size is smaller than the number of independent variables which is called high-dimensional data. The performance of two methods are compared in terms of the family wise error rate and the power of a test. Results from simulation suggest that the binary logistic regression has larger average power of a test than the gene set enrichment analysis when sample size is larger than the number of independent variables while the gene set enrichment analysis has larger average power of a test when the data is high-dimensional. However, in terms of family-wise error rate, the gene set enrichment analysis is better than the binary logistic regression in case of low-dimensional data while the binary logistic regression is superior in case of high-dimensional data.

Keywords: GSEA; binary logistic regression; LASSO; FWER; power of a test

1. บทนำ

ในยุคปัจจุบันนี้ ได้มีการนำความรู้ทางศาสตร์วิชาสถิติมาช่วยในการจัดการและวิเคราะห์ข้อมูลต่าง ๆ กันอย่างกว้างขวาง โดยเฉพาะอย่างยิ่งทางด้านการแพทย์ ซึ่งส่วนใหญ่จะเป็นการศึกษาเกี่ยวกับยีนของสิ่งมีชีวิตกับลักษณะที่ปรากฏหรือแสดงออกมาภายนอกที่เรียกว่าฟีโนไทป์ (phenotype) ซึ่งลักษณะของฟีโนไทป์ที่แสดงออกมาส่วนใหญ่จะเป็นลักษณะของการแบ่งพวกออกเป็น 2 กลุ่ม ได้แก่ เป็นโรคกับไม่เป็นโรค หรือสูงกับไม่สูง เป็นต้น โดยยีน (gene) ส่วนใหญ่ในสิ่งมีชีวิตมักจะมีความสัมพันธ์กันและทำงานร่วมกันซึ่งสามารถรวมกลุ่มของยีนที่มีความสัมพันธ์กันเรียกว่ากลุ่มยีน (gene set) โดยศึกษาความสัมพันธ์ระหว่างยีนกับฟีโนไทป์นั้น นักชีววิทยาส่วนใหญ่ไม่ต้องการดูความสัมพันธ์ของยีนแต่ละตัวกับฟีโนไทป์ที่ต้องการศึกษา แต่ต้องการศึกษาภาพรวมของทั้งกลุ่มของยีนมากกว่าว่ามีความเกี่ยวข้องกับลักษณะฟีโนไทป์ที่สนใจหรือไม่

มีงานวิจัยหนึ่งได้เสนอวิธีการสำหรับศึกษาความสัมพันธ์ระหว่างกลุ่มยีนในแต่ละกลุ่มกับลักษณะของฟีโนไทป์ที่สนใจด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (gene set enrichment analysis, GSEA) [5] โดยสำหรับการหาความสัมพันธ์ที่เกิดขึ้นนี้จะเริ่มต้นพิจารณาจากการวิเคราะห์หาความสัมพันธ์ของยีนแต่ละยีนกับลักษณะของฟีโนไทป์ก่อน [7] ซึ่งเป็นการวิเคราะห์แบบตัวแปรเดียว (univariate analysis) หลังจากนั้นถึงนำค่าของระดับความสัมพันธ์ที่คำนวณได้ในตอนแรกนี้มาใช้พิจารณาหาค่าพี (p-value) สำหรับแต่ละกลุ่มของยีนที่มีความสัมพันธ์กันในแต่ละกลุ่มว่ามีนัยสำคัญทางสถิติหรือไม่ต่อไป กล่าวคือ ถ้าค่าพีที่ได้มีนัยสำคัญทางสถิติ นั่นคือ มีจำนวนยีนเป็นจำนวนมากในกลุ่มของยีนที่มีผลต่อลักษณะของฟีโนไทป์ที่สนใจนั่นเอง

จากวิธีในการศึกษาข้างต้นนั้น จะเห็นได้ว่าใน

ขั้นตอนการวิเคราะห์ของวิธี GSEA นั้น เป็นการวิเคราะห์แบบตัวแปรเดียว คือ ไม่ได้สนใจว่ายีนทำงานร่วมกันเป็นกลุ่ม ทั้ง ๆ ที่ความเป็นจริงแล้วในการทำงานของยีนส่วนมากจะทำร่วมกันเป็นกลุ่ม มากกว่าที่จะทำงานแยกกันแบบเดี่ยว ๆ และนอกจากนี้ยังเป็นการศึกษาหาสาเหตุความสัมพันธ์ของกลุ่มยีนสำหรับในแต่ละกลุ่มเท่านั้นกับลักษณะของฟีโนไทป์ที่สนใจ กล่าวคือ เป็นการศึกษาแยกกันสำหรับแต่ละกลุ่มของยีน ทั้งนี้เพราะให้แค่เพียงความสำคัญของยีนที่มีความสัมพันธ์กันในแต่ละกลุ่มของยีนเท่านั้น โดยไม่ได้คำนึงถึงความสัมพันธ์นอกกลุ่มของแต่ละกลุ่มของยีนที่อาจเกิดขึ้นได้นั่นเอง ซึ่งในความเป็นจริงแล้วควรที่จะศึกษากลุ่มของยีนในแต่ละกลุ่มพร้อม ๆ กัน มากกว่าที่จะศึกษาทีละกลุ่มของยีนแยกกัน

วิธีการหนึ่งที่ผู้วิจัยเสนอเพื่อศึกษาความสัมพันธ์ระหว่างกลุ่มของยีนในแต่ละกลุ่มพร้อม ๆ กันกับลักษณะของฟีโนไทป์ที่สนใจ โดยที่ให้ความสนใจกับการทำงานร่วมกันของยีนเป็นกลุ่มด้วย คือ การถดถอยลอจิสติกทวิภาค (binary logistic regression) [1] ทั้งนี้เนื่องจากข้อมูลของตัวแปรตาม (ลักษณะของฟีโนไทป์ที่สนใจ) เป็นตัวแปรจำแนกประเภท (categorical variable) และแบ่งออกเป็นข้อมูลทวิภาคได้เป็น 2 ประเภท (dichotomous data or binary data) และสำหรับข้อมูลของตัวแปรอิสระ (กลุ่มของยีน) เป็นตัวแปรเชิงปริมาณ (เนื่องจากข้อมูลกลุ่มของยีนนั้นสามารถแสดงออกมาในรูปของการแสดงออกของยีน (gene expression) ซึ่งการแสดงออกของยีนสามารถถอดรหัสออกมาเป็นค่าตัวเลขได้โดยใช้เครื่องมือทางอณูชีววิทยา (molecular biology) โดยการตรวจสอบจำนวนโมเลกุลของ mRNA ซึ่งสามารถที่จะวิเคราะห์ด้วยการถดถอยลอจิสติกทวิภาคได้ นอกจากนี้ในการวิเคราะห์การถดถอยลอจิสติกทวิภาคนี้จะสามารถศึกษากลุ่มของยีนในแต่ละกลุ่มพร้อม ๆ กันได้ จึงน่าจะมีความเหมาะสมมากกว่าวิธี GSEA สำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาคนี้โดยทั่วไปแล้วจะสามารถวิเคราะห์ได้ก็ต่อเมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระที่ศึกษา $(n > p)$ เท่านั้น

ในความเป็นจริงแล้วยีนของสิ่งมีชีวิตนั้นมีอยู่เป็นจำนวนมาก และอาจมีจำนวนที่มากกว่าขนาดตัวอย่างอีกด้วย กล่าวคือ ขนาดตัวอย่างนั้นน้อยกว่าจำนวนของตัวแปรอิสระที่ศึกษา $(n < p)$ โดยข้อมูลในลักษณะนี้เราเรียกว่าข้อมูลที่มีมิติสูง (high-dimensional data) ซึ่งจากสาเหตุข้างต้นนี้เองจะเห็นได้ว่าการวิเคราะห์ด้วยการถดถอยลอจิสติกทวิภาคแบบปกติดั้งเดิมนั้นไม่สามารถทำได้ ดังนั้นวิธีการหนึ่งที่ได้รับความนิยมสำหรับจัดการและวิเคราะห์ข้อมูลที่มีมิติสูง คือ วิธีการถดถอย penalized (penalized regression) โดยวิธี LASSO [4,11] ซึ่งจะทำให้สามารถเลือกตัวแปรอิสระเข้าสู่ตัวแบบ และประมาณค่าของสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระได้ โดยจะนำตัวแปรอิสระที่ถูกเลือกเข้านั้นมาวิเคราะห์ต่อด้วยการถดถอยลอจิสติกทวิภาคแบบปกติดั้งเดิมต่อไป เพื่อหาความสัมพันธ์ระหว่างกลุ่มของยีนในแต่ละกลุ่มพร้อม ๆ กันกับลักษณะของฟีโนไทป์ที่สนใจ โดยที่อยู่บนพื้นฐานที่ว่ายีนมีการทำงานร่วมกันเป็นกลุ่ม

Subramanian และคณะ [5] ได้ศึกษาเพื่อแสดงถึงประสิทธิภาพของวิธี GSEA โดยใช้ข้อมูลที่เกี่ยวข้องกับผู้ป่วยโรคมะเร็ง (มะเร็งเม็ดเลือดขาวและมะเร็งปอด) มาวิเคราะห์ สรุปได้ว่าพบนัยสำคัญทางสถิติของกลุ่มของยีนในทุก ๆ ชุดข้อมูล นอกจากนี้ Abatangelo และคณะ [10] ได้ศึกษาเปรียบเทียบวิธีในการวิเคราะห์ความสัมพันธ์ในข้อมูลการแสดงออกของยีนโดยดูกลุ่มของยีนเป็นหลักใน 4 วิธี ได้แก่ Fisher's exact test, GSEA, RS (random-set) และ GLAPA (gene list analysis with prediction accuracy) สรุปได้ว่าในแต่ละวิธีค่อนข้างให้ผลที่มี

ความแตกต่างกันและวิธี GSEA จะมีความคงเส้นคงวา (consistent) มากที่สุด

Yusuff และคณะ [9] ได้นำตัวแบบการถดถอยลอจิสติกทวิภาคไปใช้ศึกษาเกี่ยวกับโรคมะเร็งเต้านมใน 4 ปัจจัย โดยใช้ข้อมูลที่เก็บได้จากแบบสอบถามของคนไข้จำนวน 130 ชุด สรุปได้ว่าพบนัยสำคัญทางสถิติของทั้ง 4 ปัจจัย นอกจากนี้ Mount และคณะ [8] ได้นำตัวแบบการถดถอยลอจิสติกทวิภาคไปใช้ศึกษาเกี่ยวกับการปรับปรุงการการพยากรณ์ของข้อมูลการแสดงออกของยีน สรุปได้ว่า B cell-related gene เป็นปัจจัยที่สำคัญที่สุดในการกำหนดหรือทำนายผลของคนไข้

ในการศึกษาครั้งนี้ ผู้วิจัยมีความสนใจในการเปรียบเทียบวิธีการศึกษาความสัมพันธ์ระหว่างกลุ่มของยีนและฟีโนไทป์ทวิภาคระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและการถดถอยลอจิสติกทวิภาค โดยจะหาค่าพีจากแต่ละวิธี และนำค่าพีที่ได้นี้มาใช้ในการหาค่าอัตราความผิดพลาดรวม (family wise error rate, FWER) และกำลังการทดสอบ (power of a test) เพื่อเปรียบเทียบว่าวิธีการใดใน 2 วิธี ข้างต้นที่มีประสิทธิภาพและมีความเหมาะสมในการศึกษาเรื่องนี้มากที่สุด

2. วิธีการวิจัย

2.1 ศึกษาค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง

2.1.1 การวิเคราะห์ความสำคัญของกลุ่มยีน (gene set enrichment analysis, GSEA) [5]

เป็นวิธีที่ได้รับความนิยมซึ่งใช้สำหรับวิเคราะห์นัยสำคัญทางสถิติของกลุ่มยีนที่มีความสัมพันธ์กัน ซึ่งความสัมพันธ์จะสอดคล้องตามลักษณะของความแตกต่างระหว่างลักษณะ 2 ลักษณะ ที่สนใจ เช่น ฟีโนไทป์ โดยความสัมพันธ์ของกลุ่มของยีนในที่นี้อาจเป็นความสัมพันธ์ในลักษณะของการเชื่อมโยงทางชีวภาพที่มีตั้งแต่เริ่มต้น (prior biological pathway) หรือการแสดงออกร่วมของยีน (co-expression) ในการทดลองก่อนหน้า เป็นต้น โดยมีขั้นตอนในการวิเคราะห์ 6 ขั้นตอน ดังนี้

(1) เตรียมข้อมูลของยีนทั้งหมด p ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด n ตัวอย่าง โดยแสดงได้ในรูปของ gene expression matrix ดังรูปที่ 1

		Samples			
		1	2	...	n
Genes	1	x_{11}	x_{21}	...	x_{n1}
	2	x_{12}	x_{22}	...	x_{n2}
	⋮	⋮	⋮	⋱	⋮
	p	x_{1p}	x_{2p}	...	x_{np}

รูปที่ 1 ลักษณะข้อมูลของยีนในรูปของ gene expression matrix

และแบ่งกลุ่มยีนสำหรับยีนที่มีความสัมพันธ์กัน เรียกว่ากลุ่มยีน (gene set) ซึ่งแทนด้วยสัญลักษณ์ S ตัวอย่าง เช่น Gene set 1 $(S_1) = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k\}$

(2) คำนวณหาค่าสหสัมพันธ์ระหว่างยีนแต่ละยีนทั้งหมด p ยีนที่แบ่งตามลักษณะของ ฟีโนไทป์ที่สนใจ ซึ่งในที่นี้กำหนดให้มีแค่เพียง 2 ลักษณะ นั่นคือ 0 กับ 1 กล่าวคือ เป็นการหาค่าความ สัมพันธ์ระหว่างตัวแปรตาม (y) ที่มีค่าได้เพียง 2 ค่าเท่านั้นกับตัวแปรอิสระ $(\tilde{x})$ ซึ่งเป็นตัวแปรเชิงปริมาณ (interval or ratio) ดังนั้นค่าสหสัมพันธ์ในที่นี้ คือ point biserial correlation $\left(r_{pb}\right)$ [7]

โดยที่ $r_{pb} = \dfrac{\overline{x}_1 - \overline{x}_0}{\overline{s}_x}\sqrt{p_1 p_0}$ (1)

และ $-1 \le r_{pb} \le 1$ เมื่อ $\overline{x}_1$ และ $\overline{x}_0$ แทนค่าเฉลี่ยของตัวแปรเชิงปริมาณ ในกลุ่มของลักษณะที่เป็น 1 และ 0 ตามลำดับ; p_1 และ p_0 ค่าสัดส่วนของข้อมูลตัวอย่างในกลุ่มของลักษณะที่เป็น 1 และ 0 ตามลำดับ (โดย $p_1 = \frac{n_1}{n}$ และ $p_0 = \frac{n_0}{n}$ เมื่อ $n = n_0 + n_1$ แทนจำนวนตัวแปรทั้งหมด); s_x แทนส่วนเบี่ยงเบนมาตรฐานของตัวแปรเชิงปริมาณทั้งหมด (ทั้ง 2 กลุ่ม)

(3) เรียงลำดับยีนทั้ง p ยีน ตามค่าสัมบูรณ์ของ point biserial correlation : $|r_{pb}|$ โดยเรียงจากค่า $|r_{pb}|$ ที่มากที่สุดไปยังค่า $|r_{pb}|$ ที่น้อยที่สุด จะได้เซตของลำดับยีน (rank list) ซึ่งแทนด้วยสัญลักษณ์ L นั่นคือ $L = \{x_1, x_2, \ldots, x_p\}$ โดยที่ $r_{pb}(x_i) = r_j$ เมื่อ $r_{pb}(x_j)$ แทน ค่าของ point biserial correlation ของ x_j ซึ่งเขียนในรูปแบบอย่างง่าย คือ r_j

(4) คำนวณหาค่า enrichment score (ES) ของแต่ละเซตของยีน : $ES^*(S_i)$; $i = 1, 2, \ldots, t$

โดยกำหนดให้ $P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^\alpha}{\sum_{g_i \in S} |r_j|^\alpha}$ (2)

และ $P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(p - p_s)}$ (3)

โดยที่ α แทนการถ่วงน้ำหนักของยีนในกลุ่มของยีนกับฟีโนไทป์; p แทนจำนวนของยีนทั้งหมด; p_s แทนจำนวนของยีนในกลุ่มของยีน

จะได้ว่า $ES^*(S_i) = \max\{ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i)\}$ (4) สำหรับ $i = 1, 2, \ldots, t$

(5) เรียงสับเปลี่ยน (permutation) ในส่วนของค่าแสดงลักษณะของฟีโนไทป์ $\{0, 1\}$ ทั้งหมดแล้วกลับไปเริ่มทำในขั้นตอนที่ 2 ถึงขั้นตอนที่ 5 ใหม่ โดยทำทั้งหมด m ครั้ง โดยที่ครั้งที่เรียงสับเปลี่ยน $ES^{(l)}(S_i)$ สำหรับ 1 $\Rightarrow \{ES^{(1)}(S_i)\}$; สำหรับ 2 $\Rightarrow \{ES^{(2)}(S_i)\}$; ... ; สำหรับ m $\Rightarrow \{ES^{(m)}(S_i)\}$ สำหรับทุก $i = 1, 2, \ldots, t$ และ $l = 1, 2, \ldots, m$

(6) คำนวณหาค่าพี $(p - value_i)$ สำหรับแต่ละกลุ่มของยีน (S_i) ทุก $i = 1, 2, \ldots, t$ โดยที่

สำหรับ $ES^*(S_i) > 0$ จะได้ว่า

$$p - value_i = \frac{\text{จำนวนของ } ES^{(l)}(S_i) \ge ES^*(S_i)}{m}$$

บางค่า $l = 1, 2, \ldots, m$ (5)

สำหรับ $ES^*(S_i) < 0$ จะได้ว่า

$$p - value_i = \frac{\text{จำนวนของ } ES^{(l)}(S_i) \le ES^*(S_i)}{m}$$

บางค่า $l = 1, 2, \ldots, m$ (6)

2.1.2 การวิเคราะห์การถดถอยลอจิสติก (logistic regression analysis)

เป็นการวิเคราะห์ที่มีวัตถุประสงค์เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยที่ตัวแปรตามจะเป็นตัวแปรจำแนกประเภท (categorical variable) และอาจแบ่งออกได้เป็นข้อมูลทวิภาค (dichotomous data) ซึ่งก็คือข้อมูล 2 กลุ่ม หรือมากกว่า 2 กลุ่ม สำหรับการวิเคราะห์การถดถอยลอจิสติกกรณีที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่าการวิเคราะห์การถดถอยลอจิสติกทวิภาค (binary logistic regression) โดยสำหรับการทดสอบสมมุติฐานที่ใช้ในการศึกษาครั้งนี้คือการทดสอบอัตราส่วนภาวะน่าจะเป็น (likelihood ratio test, LRT) ซึ่งเป็นการทดสอบสมมุติฐานทางสถิติของตัวแบบที่มีค่าสัมประสิทธิ์ที่ได้จากการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ เทียบกับตัวแบบที่ไม่มีค่าสัมประสิทธิ์ว่า มีความแตกต่างกันอย่างมีนัยสำคัญทางทางสถิติหรือไม่

โดยมีสมมุติฐาน คือ

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$; 1 \le k \le p$

H_a : มีอย่างน้อย 1 ค่าที่ $\beta_k \neq 0$

$; k = 1, 2, \ldots, p$

ตัวสถิติสำหรับทดสอบสมมุติฐาน คือ

$$LRT = -2\log\left(\frac{\ell_0}{\ell_1}\right) = -2\log\ell_0 + 2\log\ell_1 \sim \chi_k^2 \quad (7)$$

เมื่อ ℓ_0 คือ ค่าที่มากที่สุดของฟังก์ชันภาวะน่าจะเป็น (likelihood function) ภายใต้ H_0 หรือตัวแบบลดรูป (reduced model); ℓ_1 คือ ค่าที่มากที่สุดของฟังก์ชันภาวะน่าจะเป็น (likelihood function) ภายใต้ตัวแบบเต็ม (full model)

2.1.3 วิธีการถดถอย penalized (penalized regression)

เป็นวิธีที่ค่อนข้างเป็นที่นิยมใช้กันอย่างแพร่หลายสำหรับการศึกษาข้อมูลที่มีมิติสูงโดยในการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระ จะหาได้จากการหาค่า $\hat{\beta}$ ที่ทำให้ฟังก์ชันเป้าหมาย มีค่าต่ำสุด โดย

$$\hat{\beta} = \arg\min_{\beta} \left\| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right\|^2 + P_\lambda(\beta) \quad ; i = 1, 2, \ldots, n \quad (8)$$

เมื่อ $P_\lambda(\beta)$ คือ ฟังก์ชัน penalty (penalty function) และ λ คือ พารามิเตอร์ที่มีการปรับค่าแล้ว (tuning parameter) ซึ่ง $\lambda \ge 0$

จะเห็นได้ว่าสำหรับการหาค่า $\hat{\beta}$ สำหรับวิธีการถดถอย penalized จะมีฟังก์ชัน penalty เพิ่มขึ้นมาอีกพจน์หนึ่ง สำหรับค่าของพารามิเตอร์ที่มีการปรับค่าแล้ว โดยทั่วไปจะใช้วิธี cross-validation ในการหาค่าที่เหมาะสมสำหรับข้อมูลที่ต้องการวิเคราะห์ ทั้งนี้หากเราเลือกฟังก์ชัน penalty ที่เหมาะสมแล้วก็จะทำให้สมการข้างต้นนั้นสามารถคัดกรองตัวแปรเข้าในตัวแบบได้อีกด้วย กล่าวคือฟังก์ชัน penalty นั้นจะทำให้สัมประสิทธิ์บางตัวในการประมาณค่ามีค่าเท่ากับศูนย์นั่นเอง

ฟังก์ชัน penalty ของวิธี LASSO (least absolute shrinkage and selection operator) นำเสนอโดย Tibshirani (1996) ซึ่งมีวัตถุประสงค์เพื่อใช้เป็นทั้งวิธีที่สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระ $(\tilde{\beta})$ โดยวิธีนี้จะใช้ $\ell_1 - norm$ ในการปรับค่าด้วยวิธีกำลังสองน้อยที่สุด ซึ่งมีฟังก์ชัน penalty ดังนี้

$$P_\lambda(\beta) = \lambda\sum_{j=1}^{p}|\beta_j| \quad (9)$$

ซึ่งค่า $\hat{\beta}$ หาได้จาก

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \left\| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right\|^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad ; i = 1, 2, \ldots, n \quad (10)$$

โดยตัวประมาณที่ได้จากวิธี LASSO จะมีค่า $\hat{\beta}$ ส่วนใหญ่เป็นศูนย์ และมีค่า $\hat{\beta}$ บางส่วนไม่เท่ากับศูนย์ (sparse estimator) ดังนั้นวิธี LASSO จึงเป็นวิธีที่สามารถเลือกตัวแปรเข้าได้โดยอัตโนมัติ (ตัวแปรที่ $\hat{\beta} \neq 0$) นั่นเอง นอกจากนี้แล้วการใช้วิธี LASSO ในการวิเคราะห์ข้อมูลที่มีมิติสูงยังมีข้อจำกัดอยู่ กล่าวคือ วิธี LASSO สามารถเลือกตัวแปรอิสระเข้าได้มากที่สุดจำนวน n ตัว ซึ่งถ้าข้อมูลในการวิเคราะห์ของเรามีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเป็นจำนวนมาก ทำให้วิธี LASSO อาจไม่ค่อยเหมาะสม

ที่จะใช้ในการวิเคราะห์ และสำหรับกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง วิธี LASSO มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวเดียวจากกลุ่มของตัวแปรอิสระมีความสัมพันธ์กันสูงเข้าตัวแบบ โดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม โดยวิธี LASSO จะมีประสิทธิภาพสูงในการพยากรณ์ เมื่อตัวแบบจริงมีจำนวนของตัวแปรอิสระไม่มากนักที่มีความสัมพันธ์กับตัวแปรตามและขนาดของ $\hat{\beta}$ ที่ไม่เท่ากับศูนย์มีขนาดใหญ่ [4]

2.2 กำหนดการจำลองข้อมูล

2.2.1 กำหนดค่าเริ่มต้นและรูปแบบของตัวแบบที่ใช้จำลองข้อมูล โดยจะสร้างข้อมูลที่มีจำนวนของขนาดตัวอย่าง n ค่า และพารามิเตอร์จำนวน p ตัว ดัง 2 กรณีต่อไปนี้

(1) กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n=100,\ p=30)$

(2) กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n=100,\ p=300)$

2.2.2 กำหนดค่าสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระ โดยแบ่งเป็น 2 กรณีศึกษา ดังนี้

(1) กรณีที่ 1 : ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ทั้งหมดโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระ ในกลุ่มของ *Gene set* 1 (S_1) 5 ตัวแปรอิสระ $(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)$, *Gene set* 2 (S_2) 5 ตัวแปรอิสระ $(\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10})$ และของตัวแปรอิสระที่เหลือมีค่าเป็น 0

(2 กรณีที่ 2 : มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระ ในกลุ่มของ *Gene set* 1 (S_1) 4 ตัวแปรอิสระ $(\beta_1,\beta_2,\beta_3,\beta_4)$, *Gene set* 2 (S_2) 3 ตัวแปรอิสระ $(\beta_6,\beta_7,\beta_8)$, *Gene set* 3 (S_3) 3 ตัวแปรอิสระ $(\beta_{11},\beta_{12},\beta_{13})$ และของตัวแปรอิสระที่เหลือมีค่าเป็น 0

2.2.3 กำหนดกลุ่มตัวแปรอิสระตามลักษณะความสัมพันธ์ของยีน $\left(S_i;i=1,2,\ldots,\left(\frac{p}{5}\right)\right)$ ดังนี้ *Gene set* 1 $(S_1)=\{\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_5\}$, *Gene set* 2 $(S_2)=\{\tilde{x}_6,\tilde{x}_7,\ldots,\tilde{x}_{10}\}$, ..., *Gene set* $\left(\frac{p}{5}\right)$ $\left(S_{\left(\frac{p}{5}\right)}\right)=\left\{\tilde{x}_{\left(\frac{p}{5}-4\right)},\tilde{x}_{\left(\frac{p}{5}-3\right)},\ldots,\tilde{x}_{\left(\frac{p}{5}\right)}\right\}$

*สำหรับขอบเขตงานวิจัยนี้ได้กำหนดให้แต่ละกลุ่มของยีนประกอบด้วยยีนทั้งหมด 5 ยีน ในทุก ๆ กลุ่มยีน

2.3 จำลองข้อมูลจากค่าเริ่มต้นที่กำหนด

2.3.1 จำลองข้อมูลตัวแปรอิสระดังต่อไปนี้

(1) ตัวแปรอิสระมีการแจกแจงแบบปกติมาตรฐานหลายตัวแปร : $\tilde{x}_i \sim N_p(\tilde{0},I)$; $i=1,2,\ldots,p$

(2) ตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร โดยมีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum(p\times p)$: $\tilde{x}_i \sim N_p(\tilde{0},\sum)$; $i=1,2,\ldots,p$

โดยที่

$$\Sigma=\begin{bmatrix} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{5\times 5} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{5\times 5} \end{bmatrix}$$

(โดยกำหนดสหสัมพันธ์ระหว่างตัวแปรอิสระ $\rho=0.5$)

2.3.2 จำลองข้อมูลตัวแปรตามที่มีการแจกแจงแบบแบร์นูลลี : $y_i \sim Bernoulli(\pi_i)$ สำหรับ $i=1,2,\ldots,n$ โดยที่ $\pi_i=\dfrac{\exp(\sum_{j=1}^{p}x_{ij}\beta_j)}{1+\exp(\sum_{j=1}^{p}x_{ij}\beta_j)}$ และ $0\le\pi_i\le 1$

2.4 นำข้อมูลที่ได้จากการจำลองมาวิเคราะห์ตามขั้นตอนต่อไปนี้ สำหรับแต่ละกรณีที่ศึกษา

2.4.1 กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n = 100,\ p = 30)$

(1) ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่าพีของแต่ละกลุ่มของยีน

(1.1) วิธีการถดถอยลอจิสติกทวิภาค

(1.2) วิธี GSEA

2.4.2 กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n = 100,\ p = 300)$

(1) ใช้วิธี LASSO ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ (เฉพาะสำหรับวิธีการถดถอยลอจิสติกทวิภาค) โดยเลือกทั้งกลุ่มของยีนที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์อย่างน้อย 1 ตัว

(2) ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่าพีของแต่ละเซตของยีน

(2.1) วิธีการถดถอยลอจิสติกทวิภาค (สำหรับกลุ่มของยีนที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระเท่ากับศูนย์ทุกตัวจะถือว่าค่าพีของกลุ่มยีนนั้น ๆ มีค่าเป็น 1)

(2.2) วิธี GSEA

2.5 เปรียบเทียบประสิทธิภาพของวิธีการศึกษาความสัมพันธ์ของกลุ่มของยีนและฟีโนไทป์ทวิภาค

โดยใช้ค่าอัตราความผิดพลาดรวมและกำลังการทดสอบเป็นเกณฑ์ในการพิจารณาเปรียบเทียบ

2.5.1 อัตราความผิดพลาดรวม (family wise error rate, FWER)

เป็นโอกาสของการกระทำความผิดพลาดแบบที่ 1 (type I error, α) อย่างน้อยหนึ่งครั้งของชุดการเปรียบเทียบ หรือเป็นโอกาสของชุดการเปรียบเทียบ (set or family of contrasts) จำนวน 1 ชุด จะมีการตัดสินใจผิดพลาดแบบที่ 1 เกิดขึ้น ซึ่งความผิดพลาดแบบนี้เกิดขึ้นในการเปรียบเทียบความแตกต่างค่าเฉลี่ยจำนวนหลายค่าหรือหลายกลุ่มค่าเฉลี่ยแล้วได้ข้อสรุปของการเปรียบเทียบดังกล่าวจำนวน 1 ชุด (set or family) [3] โดยจะได้ว่าความผิดพลาดแบบที่ 1 เกิดจากการที่ปฏิเสธสมมุติฐานว่าง (null hypothesis, H_0) เมื่อสมมุติฐานว่างเป็นจริง

ซึ่งอัตราความผิดพลาดรวมสามารถคำนวณได้จาก

$$FWER = P\left(\text{เกิด } Type\ I\ error\right) = \frac{\text{จำนวนของการจำลองที่เกิดความผิดพลาดแบบที่ 1 อย่างน้อยหนึ่งครั้ง}}{\text{จำนวนของการจำลองทั้งหมด}} \quad (11)$$

(โดยในการจำลองแต่ละครั้ง เราจะปฏิเสธ H_0^C ก็ต่อเมื่อ $P_h^C < \alpha$ และค่าวัดประสิทธิภาพ FWER นี้ ยิ่งต่ำมากเท่าไรก็จะยิ่งดี) [6]

2.5.2 กำลังการทดสอบ (power of a test)

กระบวนการขั้นตอนของการทดสอบสมมุติฐานจะต้องตั้งสมมุติฐานสองแบบ กล่าวคือ สมมติฐานว่าง (null hypothesis, H_0) ซึ่งโดยทั่วไปจะเป็นสมมุติฐานที่ไม่มีการเปลี่ยนแปลงไปจากสภาวะเดิมหรือไม่มีผลที่แตกต่างจากของเดิม ส่วนอีกสมมุติฐานหนึ่ง คือ สมมุติฐานทางเลือกอื่นหรือสมมุติฐานแย้ง (alternative hypothesis, H_a) ซึ่งโดยทั่วไปจะเป็นสมมุติฐานที่เกี่ยวกับความเชื่อที่ต้องการทดสอบ โดยในการสรุปผลมักจะเกิดความผิดพลาดได้สองแบบ กล่าวคือ

ความผิดพลาดแบบที่ 1 (type I error, α) โดยที่ $\alpha = P\left(Reject\ H_0 \middle| H_0\ is\ true\right)$ และ

ความผิดพลาดแบบที่ 2 (type II error, β) โดยที่ $\beta = P\left(Accept\ H_0 \middle| H_0\ is\ false\right)$

โดยทั่วไปแล้วปัญหาในการทดสอบสมมุติฐานจะพยายามควบคุม α ให้มีค่าน้อยและจะพยายามทำให้ β มีค่าน้อยที่สุดเพื่อที่ทำให้ $1-\beta$ มีค่ามากที่สุด [2] ซึ่งเรียก $1-\beta$ ว่ากำลังการทดสอบ โดยสำหรับการทดสอบใด ๆ ที่มีกำลังการทดสอบยิ่งมากจะแสดงว่าการทดสอบนั้นยิ่งดี

ซึ่งกำลังการทดสอบสามารถคำนวณได้จาก

$$\text{power of a test} = P\left(Reject\ H_0 \middle| H_0\ is\ false\right) = \frac{\text{จำนวนครั้งของการปฏิเสธ } H_0 \text{ เมื่อ } H_0 \text{ เป็นเท็จ}}{\text{จำนวนของ } H_0 \text{ ที่เป็นเท็จทั้งหมด}} \quad (12)$$

2.6 วิเคราะห์ผลลัพธ์ที่ได้จากการเปรียบเทียบค่าอัตราความผิดพลาดรวมและกำลังการทดสอบที่ได้จากทั้งสองวิธี

โดยจำแนกตามลักษณะของข้อมูล (ขนาดตัวอย่างและจำนวนของตัวแปรอิสระ) ลักษณะความสัมพันธ์ของยีนในกลุ่มยีน และสหสัมพันธ์ของตัวแปรอิสระที่ศึกษาและสรุปผล

3. ผลการวิจัย

สำหรับงานวิจัยนี้ นำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 2 ส่วน คือ ในส่วนที่ 1 เปรียบเทียบค่าอัตราความผิดพลาดรวมจากการทดสอบสมมุติฐานระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและวิธีการถดถอยลอจิสติกทวิภาค และในส่วนที่ 2 เปรียบเทียบกำลังการทดสอบจากการทดสอบสมมุติฐานระหว่าง 2 วิธี ดังกล่าว เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในกลุ่มยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา และสหสัมพันธ์ระหว่างตัวแปรอิสระในระดับต่าง ๆ ที่สนใจศึกษา ดังนี้

จากตารางที่ 1 พบว่าเมื่อขนาดตัวอย่างเท่ากับ 100 และจำนวนของตัวแปรอิสระเท่ากับ 30 $(n > p)$ การศึกษาด้วยวิธี GSEA จะมีความเหมาะสมมากที่สุด เนื่องจากวิธี GSEA มีค่าอัตราความผิดพลาดรวมต่ำกว่าเมื่อเปรียบเทียบกับวิธี การถดถอยลอจิสติกทวิภาคในทุก ๆ สถานการณ์ที่จำลองข้อมูล และในกรณีที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาค่าอัตราความผิดพลาดรวมของทั้งสองวิธีจะมีแนวโน้มลดลง ในขณะที่เมื่อขนาดตัวอย่างเท่ากับ 100 และจำนวนของตัวแปรอิสระเท่ากับ 300 $(n < p)$ การศึกษาด้วยวิธีการถดถอยลอจิสติกทวิภาคจะมีความเหมาะสมมากที่สุด เนื่องจากวิธีการถดถอยลอจิสติกทวิภาคมีค่าอัตราความผิดพลาดรวมต่ำกว่าเมื่อเปรียบเทียบกับวิธี GSEA ในทุก ๆ สถาน-การณ์ที่จำลองข้อมูล

จากตารางที่ 2 พบว่าเมื่อขนาดตัวอย่างเท่ากับ 100 และจำนวนของตัวแปรอิสระเท่ากับ 30 $(n > p)$ การศึกษาด้วยวิธีการถดถอยลอจิสติกทวิภาคจะมีความเหมาะสมมากที่สุด เนื่องจากวิธีการถดถอยลอจิสติกทวิภาคมีกำลังการทดสอบเฉลี่ยมากกว่าเมื่อเปรียบเทียบกับวิธี GSEA ในทุก ๆ สถานการณ์ที่จำลองข้อมูล และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานซึ่งเป็นค่าที่อยู่ในวงเล็บ จะเห็นได้ว่าค่าส่วนเบี่ยงเบนมาตรฐาน

ของวิธี GSEA มีค่าค่อนข้างสูงเมื่อเปรียบเทียบกับของวิธีการถดถอยลอจิสติกทวิภาคซึ่งแสดงถึงความผันผวนที่ค่อนข้างมากของกำลังการทดสอบเฉลี่ยในวิธี GSEA ในขณะที่เมื่อขนาดตัวอย่างเท่ากับ 100 และจำนวน

ตารางที่ 1 ค่าอัตราความผิดพลาดรวม (FWER) ของแต่ละสถานการณ์จากข้อมูลจำลอง

ขนาดตัวอย่าง (n)	สหสัมพันธ์ระหว่างตัวแปรอิสระ (ρ)	ค่าอัตราความผิดพลาดรวม (FWER)			
		ลักษณะความสัมพันธ์ของยีนในกลุ่มยีน			
		ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด		มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	
		GSEA	binary logistic	GSEA	binary logistic
$n > p$	$\rho = 0.0$	**0.01***	0.58	**0.00***	0.57
	$\rho = 0.5$	**0.00***	0.11	**0.00***	0.25
$n < p$	$\rho = 0.0$	1.00	**0.08***	0.99	**0.08***
	$\rho = 0.5$	0.99	**0.00***	0.99	**0.01***

*วิธีที่เหมาะสมที่สุดในแต่ละกรณี

ตารางที่ 2 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของกำลังการทดสอบของแต่ละสถานการณ์จากข้อมูลจำลอง

ขนาดตัวอย่าง (n)	สหสัมพันธ์ระหว่างตัวแปรอิสระ (ρ)	ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของกำลังการทดสอบ			
		ลักษณะความสัมพันธ์ของยีนในกลุ่มยีน			
		ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด		มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	
		GSEA	binary logistic	GSEA	binary logistic
$n > p$	$\rho = 0.0$	0.2950 (0.2472)	**0.9400* (0.2387)**	0.1900 (0.2024)	**0.9433* (0.1898)**
	$\rho = 0.5$	0.4750 (0.1095)	**1.0000* (0.0000)**	0.1567 (0.1738)	**0.9067* (0.1958)**
$n < p$	$\rho = 0.0$	**0.9200* (0.1842)**	0.1700 (0.3350)	**0.7467* (0.2375)**	0.0867 (0.2351)
	$\rho = 0.5$	**1.0000* (0.0000)**	0.0750 (0.2500)	**0.9300* (0.1365)**	0.0200 (0.1237)

*วิธีที่เหมาะสมที่สุดในแต่ละกรณี

ของตัวแปรอิสระเท่ากับ 300 $(n < p)$ การศึกษาด้วยวิธี GSEA จะมีความเหมาะสมมากที่สุด เนื่องจากวิธี GSEA มีกำลังการทดสอบเฉลี่ยมากกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยลอจิสติกทวิภาคในทุก ๆ

สถานการณ์ที่จำลองข้อมูลและเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานซึ่งเป็นค่าที่อยู่ในวงเล็บ จะเห็นได้ว่าค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยลอจิสติกทวิภาคมีค่าค่อนข้างสูงเมื่อเปรียบเทียบกับของวิธี GSEA ซึ่งแสดงถึงความผันผวนที่ค่อนข้างมากของกำลังการทดสอบเฉลี่ยในวิธีการถดถอยลอจิสติกทวิภาค

4. อภิปรายผล

4.1 ผลจากความแตกต่างระหว่างลักษณะของข้อมูล (ขนาดตัวอย่างและจำนวนของตัวแปรอิสระ)

ผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าอัตราความผิดพลาดรวมสำหรับวิธี GSEA เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ จะมากกว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ ในขณะที่ประสิทธิภาพในการหาค่าอัตราความผิดพลาดรวมสำหรับวิธีการถดถอยลอจิสติกทวิภาค เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระจะน้อยกว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ และประสิทธิภาพในการหากำลังการทดสอบสำหรับวิธี GSEA เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระจะน้อยกว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ ในขณะที่ประสิทธิภาพในการหากำลังการทดสอบสำหรับวิธีการถดถอยลอจิสติกทวิภาคเมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระจะมากกว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ

4.2 ผลจากความแตกต่างระหว่างลักษณะความสัมพันธ์ของยีนในกลุ่มยีน

ผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าอัตราความผิดพลาดรวมสำหรับวิธี GSEA และวิธีการถดถอยลอจิสติกทวิภาค เมื่อลักษณะความสัมพันธ์ของยีนในกลุ่มยีนที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด ไม่แตกต่างกันมากอย่างเห็นได้ชัดกับเมื่อลักษณะความสัมพันธ์ของยีนในกลุ่มยีนที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา และประสิทธิภาพในการหากำลังการทดสอบสำหรับวิธี GSEA และวิธีการถดถอยลอจิสติกทวิภาคเมื่อลักษณะความสัมพันธ์ของยีนในกลุ่มยีนที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมดจะค่อนข้างมากกว่าเมื่อลักษณะความสัมพันธ์ของยีนในกลุ่มยีนที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

4.3 ผลจากความแตกต่างของสหสัมพันธ์ระหว่างตัวแปรอิสระ

ผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าอัตราความผิดพลาดรวมสำหรับวิธี GSEA และวิธีการถดถอยลอจิสติกทวิภาคเมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างน้อยกว่าเมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.5 และประสิทธิภาพในการหากำลังการทดสอบสำหรับวิธี GSEA เมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างน้อยกว่าเมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.5 และสำหรับวิธีการถดถอยลอจิสติกทวิภาคเมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างมากกว่าเมื่อสหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.5

5. สรุป

งานวิจัยนี้ได้เปรียบเทียบวิธีการศึกษาความสัมพันธ์ของกลุ่มของยีนและฟีโนไทป์ทวิภาคภายใต้ขอบเขตงานวิจัยที่กำหนดระหว่างวิธี GSEA และวิธีการถดถอยลอจิสติกทวิภาคสำหรับหาค่าพีของแต่ละกลุ่มของยีนโดยที่คำนึงถึงความสัมพันธ์และการทำงานร่วมกันเป็นกลุ่มของยีน ซึ่งจะวิเคราะห์ค่าอัตราความผิดพลาดรวมและกำลังการทดสอบ ผลการวิจัย

พบว่าสำหรับกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ วิธีการถดถอยลอจิสติกทวิภาคมีกำลังการทดสอบเฉลี่ยสูง แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวมพบว่าวิธี GSEA มีค่าต่ำ ส่วนกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระวิธี GSEA มีกำลังการทดสอบเฉลี่ยสูง แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวมพบว่าวิธีการถดถอยลอจิสติกทวิภาคมีค่าต่ำ ซึ่งจากทั้ง 2 กรณี เมื่อพิจารณาถึงทั้งสองค่าพร้อมกันพบว่าวิธีการศึกษาความสัมพันธ์ของกลุ่มของยีนและฟีโนไทป์ทวิภาควิธีใดมีกำลังการทดสอบเฉลี่ยสูงก็จะมีค่าอัตราความผิดพลาดรวมสูงด้วยเช่นกัน ดังนั้นจึงไม่สามารถสรุปได้ว่าวิธีการในการศึกษาความสัมพันธ์ของกลุ่มของยีนและฟีโนไทป์ทวิภาควิธีใดเป็นวิธีที่ดีที่สุด อย่างไรก็ตาม สำหรับการนำไปใช้งานจริงจะขึ้นอยู่กับผู้นำไปใช้ว่าจะพิจารณาให้ความสำคัญกับประสิทธิภาพของวิธีศึกษาด้วยค่าอัตราความผิดพลาดรวมหรือกำลังการทดสอบมากกว่ากัน แล้วถึงเลือกใช้วิธีการศึกษาความสัมพันธ์ของกลุ่มของยีนและฟีโนไทป์ทวิภาคที่เหมาะสม ทั้งนี้ผลสรุปที่ได้เป็นผลภายใต้ขอบเขตการวิจัยที่ศึกษา

6. ข้อเสนอแนะ

จากงานวิจัยนี้ สำหรับผู้ที่สนใจอาจนำไปศึกษาต่อได้อีกในเรื่องของ

6.1 การกำหนดความสัมพันธ์ให้แต่ละกลุ่มของยีนอาจกำหนดให้มีจำนวนยีนที่แตกต่างกันซึ่งในความเป็นจริงแล้วยีนแต่ละกลุ่มอาจมีจำนวนยีนอยู่แตกต่างกันได้

6.2 ขอบเขตในการวิจัยเรื่องของลักษณะของข้อมูล (ขนาดตัวอย่างและจำนวนของตัวแปรอิสระ) ลักษณะความสัมพันธ์ของยีนในกลุ่มยีน และสหสัมพันธ์ระหว่างตัวแปรอิสระอาจจะมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้นได้

6.3 วิธีการคัดกรองตัวแปรในความเป็นจริงแล้วยังมีอีกหลายวิธีที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการคัดกรองตัวแปรอื่น ๆ มาใช้ในการพิจารณาได้อีก

6.4 เกณฑ์ที่ใช้ในการตัดสินใจสำหรับเปรียบเทียบประสิทธิภาพของแต่ละวิธีอาจต้องมีการใช้เกณฑ์ในการตัดสินใจเพิ่มมากกว่านี้เพื่อที่จะสามารถสรุปผลให้ชัดเจนได้ว่าวิธีใดที่มีประสิทธิภาพมากที่สุดในสองวิธี

7. รายการอ้างอิง

[1] กัลยา วานิชย์บัญชา, 2552, การวิเคราะห์ข้อมูลหลายตัวแปร, สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 589 น.

[2] ธีระพร วีระถาวร, 2536, การอนุมานเชิงสถิติขั้นกลาง : โครงสร้างและความหมาย, สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, พิมพ์ครั้งที่ 2, กรุงเทพฯ, 381 น.

[3] ไพฑูรย์ สุขศรีงาม, 2557, วิธีการเปรียบเทียบความแตกต่างค่าเฉลี่ยรายคู่หลังการวิเคราะห์ความแปรปรวน, ว.มรม.(มนุษยศาสตร์และสังคมศาสตร์) 8: 23-30.

[4] วิฐรา พึ่งพาพงศ์, 2558, บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้น สำหรับข้อมูลที่มีมิติสูง, ว. วิทยาศาสตร์และเทคโนโลยี 23: 212-223.

[5] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P., 2005, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proc. Nat. Acad. Sci. U.S.A. 102: 15545-15550.

[6] Benjamini, Y. and Hochberg, Y., 1995, Controlling the false discovery rate: A

practical and powerful approach to multiple testing, J. Royal Stat. Soc. Ser. B 57: 289-300.

[7] Kornbort, D., 2005, Point Biserial Correlation, pp. 1552-1553, In Everitt, B.S. andHowell, D.C. (Eds.), Encyclopedia of Statistics in Behavioral Sciences, Vol. 3, John Wiley & Sons, Chichester.

[8] Mount, D.W., Putnam, C.W., Centouri, S.M., Manziello, A.M., Pandey, R., Garland, L.L. and Martinez, J.D., 2004, Using logistic regression to improve the prognostic value of microarray gene expression data sets: Application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma, J. BMC Med. Genom. 7: 33.

[9] Yusuff, H., Mohamad, N., Ngah, U.K. and Yahaya, A.S., 2012, Breast cancer analysis using logistic regression, IJRRAS 10: 14-22.

[10] Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T.M., Mukherjee, S. and Ancona, N., 2009, Comparative study of gene set enrichment methods, J. BMC Bioinform. 10: 275.

[11] Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, J. Royal Stat. Soc. Ser. B 58: 267-288.