การเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและวิธีการถดถอยโลจิสติกทวิภาคในการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

นายสุธิภาส สิงห์เรือง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

A COMPARISON OF GENE SET ENRICHMENT ANALYSIS AND BINARY LOGISTIC REGRESSION FOR INVESTIGATING THE RELATIONSHIP BETWEEN GENE SETS AND A BINARY PHENOTYPE

Mr. Sutipat Singruang

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2015

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและวิธีการถดถอยโลจิสติกทวิภาคในการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค
โดย	นายสุธิภาส สิงห์เรือง
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์)

.....กรรมการ

(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร. ฐิตารีย์ รุ่งรัตน์เกษม)

สุธิภาส สิงห์เรือง : การเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนและวิธีการถดถอยโลจิสติกทวิภาคในการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค (A COMPARISON OF GENE SET ENRICHMENT ANALYSIS AND BINARY LOGISTIC REGRESSION FOR INVESTIGATING THE RELATIONSHIP BETWEEN GENE SETS AND A BINARY PHENOTYPE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. วิฐรา พึ่งพาพงศ์, 86 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์ เพื่อศึกษาและเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน และการถดถอยโลจิสติกทวิภาค ในการหาค่า p-value ของแต่ละเซตยีน โดยคำนึงถึงความสัมพันธ์และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก โดยการศึกษานี้จะเปรียบเทียบประสิทธิภาพ จากการวิเคราะห์ข้อมูลจำลองทั้งในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนของยีนหรือตัวแปรอิสระ และกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ หรือที่เรียกว่า “ข้อมูลที่มีมิติสูง” ในขอบเขตการศึกษาต่างๆกัน ในงานวิจัยนี้จะเปรียบเทียบค่าอัตราความผิดพลาดรวม และค่าอำนาจในการทดสอบเพื่อวัดประสิทธิภาพจากวิธีทั้งสอง

จากการศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่าวิธีการถดถอยโลจิสติกทวิภาค มีค่าอำนาจการทดสอบ(เฉลี่ย)สูง ในกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ ในขณะที่วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าอำนาจการทดสอบ(เฉลี่ย)สูง ในกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ แต่เมื่อพิจารณาถึงการวัดประสิทธิภาพจากค่าอัตราความผิดพลาดรวม พบว่าวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าต่ำ สำหรับกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ ในขณะที่วิธีการถดถอยโลจิสติกทวิภาค มีค่าต่ำสำหรับกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2558

5781591926 : MAJOR STATISTICS
KEYWORDS: GSEA / BINARY LOGISTIC REGRESSION / LASSO / FWER / POWER OF TEST

SUTIPAT SINGRUANG: A COMPARISON OF GENE SET ENRICHMENT ANALYSIS AND BINARY LOGISTIC REGRESSION FOR INVESTIGATING THE RELATIONSHIP BETWEEN GENE SETS AND A BINARY PHENOTYPE. ADVISOR: ASST. PROF. VITARA PUNGPAPONG, Ph.D., 86 pp.

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 test.

Results from simulation suggest that the binary logistic regression has larger power 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 power 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.

Department: Statistics
Field of Study: Statistics
Academic Year: 2015

Student's Signature .....
Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความช่วยเหลือและความเอาใจใส่จาก ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่กรุณาให้คำปรึกษา อบรมสั่งสอน และให้ข้อคิดเห็นต่างๆ ตลอดจนให้ความช่วยเหลือ คำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณท่าน รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช และ อาจารย์ ดร. ฐิตารีย์ รุ่งรัตน์เกษม กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อสอบ ตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ให้โอกาสทางการศึกษา และอบรมสั่งสอนให้ความรู้ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่คอยให้กำลังใจและความห่วงใย ส่งเสริมและสนับสนุนมาโดยตลอด และขอขอบคุณเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือ ให้คำแนะนำและเป็นกำลังใจให้กับผู้วิจัยตลอดมา

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบันนี้ ได้มีการนำความรู้ทางศาสตร์วิชาสถิติมาช่วยในการจัดการและวิเคราะห์ข้อมูลต่างๆกันอย่างกว้างขวาง โดยเฉพาะอย่างยิ่งทางด้านการแพทย์ ซึ่งส่วนใหญ่จะเป็นการศึกษาเกี่ยวกับยีนของสิ่งมีชีวิต กับลักษณะที่ปรากฏหรือแสดงออกมาภายนอก ที่เราเรียกว่า ฟีโนไทป์ ซึ่งลักษณะของฟีโนไทป์ที่แสดงออกมาส่วนใหญ่ จะเป็นลักษณะของการแบ่งพวกออกเป็น 2 กลุ่ม เช่น เป็นโรค กับไม่เป็นโรค หรือ สูงกับไม่สูง เป็นต้น โดยยีนส่วนใหญ่ในสิ่งมีชีวิตมักจะมีความสัมพันธ์กันและทำงานร่วมกัน ซึ่งเราสามารถรวมกลุ่มของยีนที่มีความสัมพันธ์กัน เรียกว่า เซตของยีน (gene set) โดยในการศึกษาความสัมพันธ์ระหว่างยีนกับฟีโนไทป์นั้น นักชีววิทยาส่วนใหญ่ไม่ต้องการดูความสัมพันธ์ของยีนแต่ละตัวกับฟีโนไทป์ที่ต้องการศึกษา แต่ต้องการศึกษาภาพรวมของทั้งเซตของยีนมากกว่า ว่ามีความเกี่ยวข้องกับลักษณะฟีโนไทป์ที่สนใจหรือไม่

มีงานวิจัยหนึ่งได้เสนอวิธีการสำหรับศึกษาความสัมพันธ์ระหว่าง เซตของยีนในแต่ละเซต กับลักษณะของฟีโนไทป์ที่เราสนใจ ด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA) (Subramanian et al., 2005) โดยสำหรับการหาความสัมพันธ์ที่เกิดขึ้นนี้จะเริ่มต้นพิจารณาจากการวิเคราะห์หาความสัมพันธ์ของยีนแต่ละยีนกับลักษณะของฟีโนไทป์ก่อน ซึ่งเป็นการวิเคราะห์แบบตัวแปรเดียว (Univariate Analysis) แล้วหลังจากนั้น ถึงนำค่าของระดับความสัมพันธ์ที่คำนวณได้ในตอนแรกนี้ มาใช้พิจารณาหาค่า p - value สำหรับแต่ละเซตของยีนที่มีความสัมพันธ์กันในแต่ละเซต ว่ามีนัยสำคัญทางสถิติหรือไม่ ต่อไป กล่าวคือ ถ้าค่า p - value ที่ได้มีนัยสำคัญทางสถิติ นั่นคือ มีจำนวนยีนเป็นจำนวนมากในเซตของยีนที่มีผลต่อลักษณะของฟีโนไทป์ที่เราสนใจนั่นเอง

จากวิธีในการศึกษาข้างต้นนั้น จะเห็นได้ว่าในขั้นตอนการวิเคราะห์ของวิธี GSEA นั้นเป็นการวิเคราะห์แบบตัวแปรเดียว กล่าวคือ ไม่ได้สนใจว่ายีนทำงานร่วมกันเป็นเซต ทั้งๆที่ความเป็นจริงแล้วในการทำงานของยีนส่วนมากจะทำร่วมกันเป็นเซต มากกว่าที่จะทำงานแยกกันแบบเดี่ยวๆ และนอกจากนี้แล้วยังเป็นการศึกษาหาสาเหตุความสัมพันธ์ของเซตของยีนสำหรับในแต่ละเซตเท่านั้น กับลักษณะของฟีโนไทป์ที่เราสนใจ กล่าวคือ เป็นการศึกษาแยกกันสำหรับแต่ละเซตของยีน ทั้งนี้เพราะ

ให้แค่เพียงความสำคัญของยีนที่มีความสัมพันธ์กันในแต่ละเซตของยีนเท่านั้น โดยไม่ได้คำนึงถึงความสัมพันธ์นอกกลุ่มของแต่ละเซตของยีนที่อาจจะเกิดขึ้นได้นั่นเอง ซึ่งในความเป็นจริงแล้วควรที่จะทำการศึกษาเซตของยีนในแต่ละเซตพร้อมๆกัน มากกว่าที่จะทำการศึกษาทีละเซตของยีนแยกกัน

วิธีการหนึ่งที่ผู้วิจัยเสนอเพื่อศึกษาความสัมพันธ์ระหว่าง เซตของยีนในแต่ละเซตพร้อมๆกันกับลักษณะของฟีโนไทป์ที่เราสนใจ โดยที่ให้ความสนใจกับการทำงานร่วมกันของยีนเป็นเซตด้วย คือ การวิเคราะห์การถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) ทั้งนี้เนื่องจากข้อมูลของตัวแปรตาม(ลักษณะของฟีโนไทป์ที่เราสนใจ) เป็นตัวแปรเชิงกลุ่มที่แบ่งออกได้เป็น 2 กลุ่ม (Dichotomous Data / Binary Data) และสำหรับข้อมูลของตัวแปรอิสระ(เซตของยีน) เป็นตัวแปรเชิงปริมาณ[1] ซึ่งสามารถที่จะทำการวิเคราะห์ด้วยการถดถอยโลจิสติกทวิภาคได้ นอกจากนี้คือในการวิเคราะห์การถดถอยโลจิสติกทวิภาคนี้จะสามารถศึกษาเซตของยีนในแต่ละเซตพร้อมๆกันได้ จึงน่าจะมีความเหมาะสมมากกว่าวิธี GSEA โดยสำหรับการวิเคราะห์การถดถอยโลจิสติกทวิภาคนี้โดยทั่วไปแล้วจะสามารถทำการวิเคราะห์ได้ก็ต่อเมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระที่ทำการศึกษา $(n > p)$ เท่านั้น

ในความเป็นจริงแล้วยีนของสิ่งมีชีวิตนั้นมีอยู่เป็นจำนวนมาก และอาจจะมีจำนวนที่มากกว่าขนาดตัวอย่างอีกด้วย กล่าวคือ ขนาดตัวอย่างนั้นน้อยกว่าจำนวนของตัวแปรอิสระที่ทำการศึกษา $(n < p)$ โดยข้อมูลในลักษณะนี้เราเรียกว่า ข้อมูลที่มีมิติสูง (High-Dimensional Data) ซึ่งจากสาเหตุข้างต้นนี้เองจะเห็นได้ว่า การวิเคราะห์ด้วยการถดถอยโลจิสติกทวิภาคแบบปกติดั้งเดิมนั้นไม่สามารถทำได้ ดังนั้นวิธีการหนึ่งที่ได้รับความนิยมสำหรับจัดการและวิเคราะห์ข้อมูลที่มีมิติสูงคือ Penalized Regression โดยวิธี Lasso Tibshirani (Tibshirani, 1996) ซึ่งจะทำให้เราสามารถเลือกตัวแปรอิสระเข้าสู่ตัวแบบ และประมาณค่าของสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระได้ โดยเราจะนำตัวแปรอิสระที่ถูกเลือกเข้านั้น มาทำการวิเคราะห์ต่อด้วยการถดถอยโลจิสติกทวิภาคแบบปกติดั้งเดิมต่อไป เพื่อหาความสัมพันธ์ระหว่างเซตของยีนในแต่ละเซตพร้อมๆกันกับลักษณะของฟีโนไทป์ที่เราสนใจ โดยที่อยู่บนพื้นฐานที่ว่า ยีนมีการทำงานร่วมกันเป็นเซต

---

[1] *ข้อมูลเซตของยีน จะวัดค่าออกมาเป็นเชิงปริมาณ ทั้งนี้เนื่องจากข้อมูลเซตของยีนนั้นสามารถแสดงออกมาในรูปของการแสดงออกของยีน (Gene Expression) ซึ่ง การแสดงออกของยีนสามารถถอดรหัสออกมาเป็นค่าตัวเลขได้โดยใช้เครื่องมือทางอนูชีววิทยา โดยการตรวจสอบจำนวนโมเลกุลของ mRNA ซึ่งแสดงถึงจำนวนของ Gene Expression ในเชิงปริมาณนั่นเอง*

ในการศึกษาครั้งนี้ ผู้วิจัยมีความสนใจในการเปรียบเทียบวิธีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาคระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน และการวิเคราะห์การถดถอยโลจิสติกทวิภาค โดยจะทำการหาค่า p-value จากแต่ละวิธี และนำค่า p-value ที่ได้นี้มาใช้ในการหาค่าอัตราความผิดพลาดรวม (Family Wise Error Rate : FWER) และค่าอำนาจการทดสอบ (Power of Test) เพื่อเปรียบเทียบว่าวิธีการใดในสองวิธีข้างต้นที่มีประสิทธิภาพและมีความเหมาะสมในการศึกษาเรื่องนี้มากที่สุด

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน(Gene Set Enrichment Analysis : GSEA) และการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) ในการหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีน

1.3 ข้อตกลงเบื้องต้น

ในการศึกษาครั้งนี้จะทำการเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA) และการถดถอยโลจิสติกทวิภาคในการหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีน โดยที่กำหนดให้ข้อมูลจากการจำลองและการจัดกลุ่มความสัมพันธ์เป็นดังต่อไปนี้

- $\tilde{y} = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n \times 1$ ที่เป็นตัวแปรสุ่ม

  โดยที่ $y_i \sim Bernoulli(\pi_i)$; $i = 1, 2, \ldots, n$ และ $y_i = \begin{cases} 1 & ; \text{ เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ} \end{cases}$

  เมื่อ n แทนจำนวนรายการของข้อมูล

- $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \left[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p\right]$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด

$n \times p$ และ $\tilde{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})' \sim N_p(\tilde{0}, I)$ มีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution) โดยมีฟังก์ชันความหนาแน่นของ $\tilde{x}_i$ เป็นดังนี้

$$f(\tilde{x}_i) = \frac{1}{(2\pi)^{-\frac{p}{2}}} \exp\left\{-\frac{1}{2}\tilde{x}_i'\tilde{x}_i\right\} \qquad (1\text{-}1)$$

สำหรับ $i = 1, 2, \ldots, p$ เมื่อ p แทนจำนวนของตัวแปรอิสระ

และสำหรับ $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p]$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times p$ และ $\tilde{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})' \sim N_p(\tilde{0}, \Sigma)$ มีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม Σ ขนาด $p \times p$ โดยมีฟังก์ชันความหนาแน่นของ $\tilde{x}_i$ เป็นดังนี้

$$f(\tilde{x}_i) = \frac{1}{(2\pi)^{-\frac{p}{2}}(\det \Sigma)^{-\frac{1}{2}}} \exp\left\{-\frac{1}{2}\tilde{x}_i' \Sigma^{-1} \tilde{x}_i\right\} \qquad (1\text{-}2)$$

สำหรับ $i = 1, 2, \ldots, p$ เมื่อ p แทนจำนวนของตัวแปรอิสระ

- กำหนดกลุ่มของข้อมูลตัวแปรอิสระตามลักษณะความสัมพันธ์ของยีน โดยที่กำหนดให้แต่ละเซตของยีนประกอบด้วย ยีนทั้งหมด 5 ยีน ในทุกๆเซตของยีน กล่าวคือ ให้แต่ละเซตของยีนมีจำนวนยีนที่เท่ากันและเท่ากับ 5 ยีนในทุกๆเซตของยีน

1.4 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาในส่วนของข้อมูลจำลองใน 2 กรณี ภายใต้ขอบเขตการวิจัยดังต่อไปนี้

1.4.1 ข้อมูลจำลอง : กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$

1. ทำการจำลอง (simulate) ข้อมูลตัวแปรอิสระ

- ตัวแปรอิสระ (Independent Variables : $\tilde{x}_i$) : ยีน (genes) ทั้งหมด 30 ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด 100 ตัวอย่าง $(p=30, n=100)$ โดยกำหนดให้ข้อมูลตัวอย่างของแต่ละยีนทั้ง 30 ยีน มีการแจกแจงดังต่อไปนี้ :

  กำหนด $X=[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p]$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times p$

  เมื่อ $\tilde{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})'$

  กรณีที่ 1 : ตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution)

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

  กรณีที่ 2 : ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

2. กำหนดกลุ่มของข้อมูลตามลักษณะความสัมพันธ์ของยีน $(Gene\ Set: S_i\ ; i=1,2,\ldots,6)$ ดังนี้

$$\left.\begin{aligned} Gene\ Set\ 1\ (S_1) &= \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_5\} \\ Gene\ Set\ 2\ (S_2) &= \{\tilde{x}_6, \tilde{x}_7, \ldots, \tilde{x}_{10}\} \\ Gene\ Set\ 3\ (S_3) &= \{\tilde{x}_{11}, \tilde{x}_{12}, \ldots, \tilde{x}_{15}\} \\ &\vdots \\ Gene\ Set\ 6\ (S_6) &= \{\tilde{x}_{26}, \tilde{x}_{27}, \ldots, \tilde{x}_{30}\} \end{aligned}\right\}$$

แต่ละเซตของยีนประกอบด้วย ยีนทั้งหมด 5 ยีน ในทุกๆเซตของยีน

3. ทำการกำหนดค่าสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระ โดยแบ่งเป็น 2 กรณีศึกษา ดังนี้

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

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

4. ทำการจำลอง (simulate) ข้อมูลตัวแปรตาม

- ตัวแปรตาม (Dependent Variables : $\tilde{y}$) : ลักษณะที่สนใจ (phenotype class)

ซึ่งแบ่งออกเป็น 2 ลักษณะกล่าวคือ $y_i = \begin{cases} 1 & ; \text{เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ} \end{cases}$

ซึ่ง $y_i \sim Bernoulli(\pi_i)\ ; i=1,2,\ldots,100$ เมื่อ $\pi_i = \dfrac{\exp(\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\sum_{j=1}^{p}\beta_j x_{ij})}$

โดยที่ π_i คือ โอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และ $0 \le \pi_i \le 1$

1.4.2 ข้อมูลจำลอง : กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

1. ทำการจำลอง (simulate) ข้อมูลตัวแปรอิสระ

   - ตัวแปรอิสระ (Independent Variables : $\tilde{x}_i$) : ยีน (genes) ทั้งหมด 300 ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด 100 ตัวอย่าง $(p = 300, n = 100)$ โดยกำหนดให้ข้อมูลตัวอย่างของแต่ละยีนทั้ง 300 ยีน มีการแจกแจงดังต่อไปนี้ :

     กรณีที่ 1 : ตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution)

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

     กรณีที่ 2 : ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

     โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

     (โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

2. กำหนดกลุ่มของข้อมูลตามลักษณะความสัมพันธ์ของยีน $\left(Gene\ Set: S_i\ ; i=1,2,\ldots,60\right)$ ดังนี้

$$\left.\begin{aligned} Gene\ Set\ 1\ \left(S_1\right) &= \left\{\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_5\right\} \\ Gene\ Set\ 2\ \left(S_2\right) &= \left\{\tilde{x}_6,\tilde{x}_7,\ldots,\tilde{x}_{10}\right\} \\ &\vdots \\ Gene\ Set\ 60\ \left(S_{60}\right) &= \left\{\tilde{x}_{296},\tilde{x}_{297},\ldots,\tilde{x}_{300}\right\} \end{aligned}\right\}$$

แต่ละเซตของยีนประกอบด้วย ยีนทั้งหมด 5 ยีน ในทุกๆเซตของยีน

3. ทำการกำหนดค่าสัมประสิทธิ์ $\left(\tilde{\beta}\right)$ ของตัวแปรอิสระ โดยแบ่งเป็น 2 กรณีศึกษา ดังนี้

กรณีที่ 1 : ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ทั้งหมดโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระในกลุ่มของ

$Gene\ Set\ 1\ \left(S_1\right)$ 5 ตัวแปรอิสระ $\left(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5\right)$,

$Gene\ Set\ 2\ \left(S_2\right)$ 5 ตัวแปรอิสระ $\left(\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10}\right)$

และของตัวแปรอิสระที่เหลือมีค่าเป็น 0

กรณีที่ 2 : มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระ ในกลุ่มของ

$Gene\ Set\ 1\ \left(S_1\right)$ 4 ตัวแปรอิสระ $\left(\beta_1,\beta_2,\beta_3,\beta_4\right)$,

$Gene\ Set\ 2\ \left(S_2\right)$ 3 ตัวแปรอิสระ $\left(\beta_6,\beta_7,\beta_8\right)$,

$Gene\ Set\ 3\ \left(S_3\right)$ 3 ตัวแปรอิสระ $\left(\beta_{11},\beta_{12},\beta_{13}\right)$

และของตัวแปรอิสระที่เหลือมีค่าเป็น 0

4. ทำการจำลอง (simulate) ข้อมูลตัวแปรตาม

- ตัวแปรตาม (Dependent Variables : y_i) : ลักษณะที่สนใจ (phenotype class)

ซึ่งแบ่งออกเป็น 2 ลักษณะกล่าวคือ $y_i=\begin{cases}1 & ; \text{ เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ}\end{cases}$

ซึ่ง $y_i \sim Bernoulli\left(\pi_i\right)\ ; i=1,2,\ldots,100$ เมื่อ $\pi_i=\dfrac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}$

โดยที่ π_i คือ โอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และ $0\le\pi_i\le 1$

1.5 คำจำกัดความที่ใช้ในงานวิจัย

ข้อมูลทวิภาค (Binary Data / Dichotomous Data)

คือ ข้อมูลของตัวแปรที่มีค่าที่เป็นไปได้เพียง 2 ค่า เช่น สำเร็จ -ไม่สำเร็จ, เปิด - ปิด, ใช่ – ไม่ใช่, พอใจ – ไม่พอใจ เป็นต้น

ข้อมูลที่มีมิติสูง (High Dimensional Data)

คือ ข้อมูลสำหรับการศึกษาที่มีจำนวนตัวแปรอิสระมากกว่าจำนวนของขนาดตัวอย่าง $(p > n)$

ค่าพี (P-Value)

คือ ค่านัยสำคัญน้อยที่สุดที่ทำให้ปฏิเสธสมมติฐานว่าง นั่นคือ ค่าความน่าจะเป็นที่แสดงถึงความเสี่ยงในการปฏิเสธสมมติฐานว่าง เมื่อสมมติฐานว่างเป็นจริง หรือความน่าจะเป็นของความคลาดเคลื่อนที่เกิดจากการใช้ตัวอย่างสุ่มชุดหนึ่งเพื่อใช้ในการตัดสินใจ หากความน่าจะเป็นของความคลาดเคลื่อนมีค่าน้อยกว่าความน่าจะเป็นของความคลาดเคลื่อนที่กำหนด จะปฏิเสธสมมติฐานว่าง หรืออาจกล่าวได้ว่า ถ้ากำหนดค่าความน่าจะเป็นของความคลาดเคลื่อนที่ยอมรับได้ แต่หลักฐานจากตัวอย่างให้ค่าความน่าจะเป็นของความคลาดเคลื่อนน้อยกว่าที่กำหนด จะสามารถปฏิเสธสมมติฐานว่างได้

ยีน (Gene)

คือ หน่วยควบคุมลักษณะทางพันธุกรรมของสิ่งมีชีวิต ซึ่งยีนเป็นสารเคมีจำพวกกรดนิวคลีอิก และเป็นส่วนหนึ่งของ DNA ที่สามารถควบคุมการแสดงออกได้

เซตของยีน (Gene Set)

คือ กลุ่มยีนสำหรับยีนที่มีความสัมพันธ์กัน โดยความสัมพันธ์ที่เกิดขึ้นอาจจะเป็นความสัมพันธ์ในลักษณะของการเชื่อมโยงทางชีวภาพที่มีมาตั้งแต่เริ่มต้น หรือการแสดงออกร่วมของยีนในการทดลองก่อนหน้า

ฟีโนไทป์ (Phenotype)

คือ ลักษณะที่ปรากฏออกมา หรือลักษณะที่แสดงออกมาให้เห็นภายนอก ซึ่งเป็นผลมาจากยีน เช่น สูง มีหนังตาสองชั้น มีติ่งหู ห่อลิ้นได้ เป็นต้น

1.6 ข้อจำกัดสำหรับงานวิจัย

เนื่องจากในงานวิจัยนี้ ผู้วิจัยได้กำหนดให้แต่ละเซตของยีนมีจำนวนยีนเท่ากันในทุกๆเซตของยีน ซึ่งในความเป็นจริงยีนเซตแต่ละเซตอาจมีจำนวนยีนอยู่แตกต่างกันได้ อย่างไรก็ดี ในกรณีที่จำนวนของยีนในแต่ละเซตของยีนไม่เท่ากัน เรายังสามารถวิเคราะห์ผลได้ในทิศทางเดียวกัน กับงานวิจัยนี้ เพียงแต่จะต้องมีการปรับเปลี่ยนค่า p-value ที่ได้ ให้แปรผันตามและสอดคล้องกับจำนวนของยีนในแต่ละเซตด้วย

1.7 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีในการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาควิธีใดเหมาะสมและมีประสิทธิภาพสูงสุดสำหรับการหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีน ซึ่งจะพิจารณาจาก ค่าอัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test) ของข้อมูลที่จำลองขึ้นมาสำหรับแต่ละกรณีที่ทำการศึกษา โดยสมมติฐานที่ใช้ทดสอบคือ

$$\begin{aligned} H_0 &: \tilde{\beta}_j = 0 \\ H_a &: \tilde{\beta}_j \neq 0 \end{aligned} \quad \text{เมื่อ} \quad \tilde{\beta}_j = \left(\beta_{5j-4},\ \beta_{5j-3},\ \ldots,\ \beta_{5j}\right)' \ ; j = 1,2,\ldots,\left(p/5\right)$$

1 อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER)

เป็นโอกาสของการกระทำความผิดพลาดประเภทที่ 1 (Type I Error Rate : α) อย่างน้อยหนึ่งครั้งของชุดการเปรียบเทียบ หรือเป็นโอกาสของชุดการเปรียบเทียบ (set or family of contrasts) จำนวน 1 ชุด จะมีการตัดสินใจผิดพลาดประเภทที่ 1 เกิดขึ้น ซึ่งความผิดพลาดแบบนี้เกิดขึ้นในการเปรียบเทียบความแตกต่างค่าเฉลี่ยจำนวนหลายค่าหรือหลายกลุ่มค่าเฉลี่ย แล้วได้ข้อสรุปของการเปรียบเทียบดังกล่าวจำนวน 1 ชุด (set / family) (ไพฑูรย์ สุขศรีงาม 2557) โดยจะได้ว่าความผิดพลาดประเภทที่ 1 เกิดจากการที่ปฏิเสธสมมติฐานว่าง (Null Hypothesis : H_0) เมื่อสมมติฐานว่างเป็นจริง ซึ่งอัตราความผิดพลาดรวมสามารถคำนวณได้จาก

$$\begin{aligned} FWER &= P\left(\text{เกิด } Type\ I\ Error\right) \\ &= \frac{\text{จำนวนของการจำลองที่เกิดความผิดพลาดประเภทที่ 1 อย่างน้อยหนึ่งครั้ง}}{\text{จำนวนของการจำลองทั้งหมด}} \end{aligned} \qquad (1\text{-}3)$$

(โดยในการทำการจำลองแต่ละครั้ง เราจะทำการปฏิเสธ H_0 ก็ต่อเมื่อ $P_h < \alpha$ และค่าวัดประสิทธิภาพ FWER นี้ ยิ่งต่ำมากเท่าไร ก็จะยิ่งดี) (Benjamini, 1995)

2 ค่าอำนาจในการทดสอบ (Power of Test)

ในกระบวนการขั้นตอนของการทดสอบสมมติฐานจะต้องตั้งสมมติฐานสองแบบ กล่าวคือ สมมติฐานว่าง (Null Hypothesis $: H_0$) ซึ่งโดยทั่วไปจะเป็นสมมติฐานที่ไม่มีการเปลี่ยนแปลงไปจากสภาวะเดิมหรือไม่มีผลที่แตกต่างจากของเดิม ส่วนอีกสมมติฐานหนึ่งคือ สมมติฐานทางเลือกอื่นหรือสมมติฐานแย้ง (Alternative Hypothesis $: H_a$) ซึ่งโดยทั่วไปจะเป็นสมมติฐานที่เกี่ยวกับความเชื่อที่ต้องการทดสอบ โดยในการสรุปผลมักจะเกิดความผิดพลาดได้สองแบบ กล่าวคือ

ความผิดพลาดประเภทที่ 1 (Type I Error) : α โดยที่ $\alpha = P(Reject\ H_0 | H_0\ is\ true)$ และ
ความผิดพลาดประเภทที่ 2 (Type II Error) : β โดยที่ $\beta = P(Accept\ H_0 | H_0\ is\ false)$

โดยทั่วไปแล้วปัญหาในการทดสอบสมมติฐานจะพยายามควบคุม α ให้มีค่าน้อย และจะพยายามทำให้ β มีค่าน้อยที่สุดเพื่อที่ทำให้ $1-\beta$ มีค่ามากที่สุด (ธีระพร วีระถาวร, 2536) ซึ่งเราเรียก $1-\beta$ ว่า อำนาจการทดสอบ โดยสำหรับการทดสอบใดๆที่มีค่าอำนาจการทดสอบยิ่งมาก จะแสดงว่าการทดสอบนั้นยิ่งดี ซึ่งค่าอำนาจการทดสอบสามารถคำนวณได้จาก

$$\text{POWER} = P(Reject\ H_0 | H_0\ is\ false) = \frac{\text{จำนวนครั้งของการปฏิเสธ } H_0 \text{ เมื่อ } H_0 \text{ เป็นเท็จ}}{\text{จำนวนของ } H_0 \text{ ที่เป็นเท็จทั้งหมด}} \quad (1\text{-}4)$$

1.8 วิธีการศึกษา

1. ศึกษาค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง
2. กำหนดค่าเริ่มต้น สำหรับการจำลองข้อมูล สำหรับแต่ละกรณีที่ทำการศึกษา
   2.1 กำหนดขนาดตัวอย่าง n
   2.2 กำหนดจำนวนตัวแปรอิสระ p ตัว
   2.3 กำหนดค่าสัมประสิทธิ์การถดถอยเริ่มต้น $(\tilde{\beta})$ สำหรับแต่ละกรณีที่ทำการศึกษา
   2.4 กำหนดระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระ (ρ)
3. จำลองข้อมูลจากค่าเริ่มต้นที่กำหนด

3.1 จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติมาตรฐานหลายตัวแปร

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

3.2 จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติหลายตัวแปร โดยมีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

3.3 จำลองข้อมูลของตัวแปรตามที่มีการแจกแจงแบบแบร์นูลลี

$$y_i \sim Bernoulli(\pi_i) \quad ;i = 1, 2, \ldots, n \quad \text{โดยที่} \quad \pi_i = \frac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij})}$$

4. นำข้อมูลที่ได้จากการจำลองมาทำการวิเคราะห์ตามขั้นตอนต่อไปนี้ สำหรับแต่ละกรณีที่ทำการศึกษา

4.1 กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$

4.1.1 ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่า p-value ของแต่ละเซตของยีน

4.1.1.1 วิธี การถดถอยโลจิสติกทวิภาค

4.1.1.2 วิธี GSEA

4.2 กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

4.2.1 ใช้วิธี Lasso ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ (*เฉพาะสำหรับวิธีการถดถอยโลจิสติกทวิภาค*) โดยเลือกทั้งเซตของยีน ที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์ อย่างน้อย 1 ตัว

4.2.2 ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่า p-value ของแต่ละเซตของยีน

4.2.2.1 วิธี การถดถอยโลจิสติกทวิภาค (*สำหรับเซตของยีนที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระเท่ากับศูนย์ทุกตัว จะถือว่าค่า p-value ของเซตของยีนนั้นๆมีค่าเป็น 1*)

4.2.2.2 วิธี GSEA

5. นำข้อมูลที่ได้จากข้อ 4. มาคำนวณหาค่า อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test)

6. วิเคราะห์ผลลัพธ์โดยทำการเปรียบเทียบค่าอัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test) ที่ได้จากทั้งสองวิธี และสรุปผล

1.9 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้วิธีการวิเคราะห์สำหรับการศึกษาหาสาเหตุความสัมพันธ์ระหว่างเซตของยีนสำหรับในแต่ละเซตยีน กับลักษณะของฟีโนไทป์แบบทวิภาคที่สนใจ

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ในการศึกษาความสัมพันธ์ระหว่าง เซตของยีนในแต่ละยีนเซต กับลักษณะของฟีโนไทป์ที่เราสนใจ โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีน สามารถศึกษาโดยใช้วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (GSEA) ซึ่งเป็นวิธีที่ค่อนข้างได้รับความนิยมสำหรับการศึกษาในกลุ่มนี้ นอกจากนี้แล้ว ถ้าพิจารณาถึงลักษณะของข้อมูลที่ทำการศึกษา พบว่าข้อมูลของตัวแปรตาม ซึ่งคือ ลักษณะของฟีโนไทป์ที่เราสนใจ เป็นตัวแปรเชิงกลุ่มที่แบ่งออกได้เป็น 2 กลุ่ม (Dichotomous Data / Binary Data) และข้อมูลของตัวแปรอิสระ ซึ่งคือ เซตของยีน เป็นตัวแปรเชิงปริมาณ นั้นยังสามารถทำการศึกษาได้ด้วยวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) ได้อีกวิธีหนึ่ง ซึ่งสามารถศึกษาเซตของยีนในแต่ละเซต กับลักษณะของฟีโนไทป์ที่สนใจพร้อมๆกันได้ โดยการวิเคราะห์จะทำได้ในกรณีที่ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระที่ทำการศึกษา $(n > p)$ เท่านั้น แต่ในความเป็นจริงเราอาจจะต้องเจอกับกรณีที่ขนาดตัวอย่างนั้นน้อยกว่าจำนวนของตัวแปรอิสระที่ทำการศึกษา $(n < p)$ อย่างที่หลีกเลี่ยงไม่ได้ ซึ่งข้อมูลในลักษณะนี้เราเรียกว่า “ข้อมูลที่มีมิติสูง (High-Dimensional Data)” ทำให้ไม่สามารถใช้การวิเคราะห์การถดถอยโลจิสติกทวิภาคแบบปกติดั้งเดิมได้ ดังนั้นจำเป็นต้องใช้ Penalized Regression โดยวิธี Lasso เพื่อเป็นการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ โดยเราจะนำตัวแปรอิสระที่ถูกเลือกเข้านั้น มาทำการวิเคราะห์ต่อด้วยการถดถอยโลจิสติกทวิภาคแบบปกติดั้งเดิมต่อไป ดังนั้นในงานวิจัยนี้จะกล่าวถึง วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (GSEA) และวิธีการถดถอยโลจิสติกทวิภาค รวมไปถึงวิธีการคัดกรองตัวแปรสำหรับกรณีข้อมูลที่มีมิติสูง คือ Penalized Regression ของวิธี Lasso และเกณฑ์ที่ใช้ในการตัดสินใจเพื่อวิเคราะห์ข้อมูลและสถิติที่ได้ คือ อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test)

2.1 วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA)

เป็นวิธีการหนึ่งที่ได้รับความนิยม ซึ่งใช้สำหรับค้นหาหรือวิเคราะห์นัยสำคัญทางสถิติของกลุ่มของยีน หรือเซตของยีนที่มีความสัมพันธ์กัน (gene set) ซึ่งความสัมพันธ์จะสอดคล้องตามลักษณะ

ของความแตกต่างระหว่างลักษณะ 2 ลักษณะที่สนใจ เช่น ฟีโนไทป์ (phenotype) เป็นต้น โดยความสัมพันธ์ของเซตของยีนในที่นี้อาจจะเป็นความสัมพันธ์ในลักษณะของการเชื่อมโยงทางชีวภาพที่มีมาตั้งแต่เริ่มต้น (prior biological pathways) หรือ การแสดงออกร่วมของยีน (co-expression) ในการทดลองก่อนหน้า เป็นต้น โดยมีขั้นตอนในการวิเคราะห์ 6 ขั้นตอน ดังนี้

1. เตรียมข้อมูลของยีน ทั้งหมด p ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด n ตัวอย่าง โดยแสดงได้ในรูปของ Gene Expression Matrix ดังภาพที่ 2.1

		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}

ภาพที่ 2.1 แสดงลักษณะข้อมูลของยีนในรูปของ Gene Expression Matrix

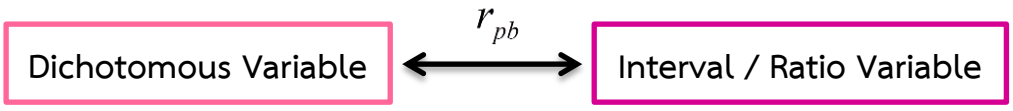
และแบ่งกลุ่มยีนสำหรับยีนที่มีความสัมพันธ์กัน เรียกว่า เซตของยีน (gene set) ซึ่งแทนด้วยสัญลักษณ์ S โดยจะเป็นความสัมพันธ์ในลักษณะของการเชื่อมโยงทางชีวภาพที่มีมาตั้งแต่เริ่มต้น หรือการแสดงออกร่วมของยีนในการทดลองก่อนหน้า ตัวอย่างเช่น

$$\text{Gene Set 1 } (S_1) = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k\} \quad ;1 \le k \le p$$

โดยที่ $\tilde{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix} ; i = 1, 2, \ldots, p$ แทน ชุดข้อมูลของยีนที่ i

2. คำนวณหาค่าสหสัมพันธ์ (correlation) ระหว่างยีนแต่ละยีน ทั้งหมด p ยีนที่แบ่งตามลักษณะของฟีโนไทป์ที่สนใจ ซึ่งในที่นี้กำหนดให้มีแค่เพียง 2 ลักษณะ/ค่า นั่นคือ 0 กับ 1 ซึ่งเขียนแทนด้วย $y = \{0, 1\}$ กล่าวคือ เป็นการหาค่าความสัมพันธ์ระหว่างตัวแปรตาม (y) กับ ตัวแปรอิสระ $(\tilde{x})$

เนื่องจาก ต้องการหาค่าสหสัมพันธ์ ระหว่าง ตัวแปรที่มีค่าได้เพียง 2 ค่าเท่านั้น คือ $\{0,1\}$ (Dichotomous Variable) กับ ตัวแปรเชิงปริมาณ (interval / ratio) ดังนั้น ค่าสหสัมพันธ์ในที่นี้ คือ Point Biserial Correlation (r_{pb}) (Howell, 2005)



โดยที่ $$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{\bar{s}_x}\sqrt{p_1 p_0} \qquad (2\text{-}1)$$

และ $-1 \le r_{pb} \le 1$

เมื่อ $\bar{x}_1$ แทน ค่าเฉลี่ยของตัวแปรเชิงปริมาณ ในกลุ่มของลักษณะที่เป็น 1

$\bar{x}_0$ แทน ค่าเฉลี่ยของตัวแปรเชิงปริมาณ ในกลุ่มของลักษณะที่เป็น 0

p_1 แทน ค่าสัดส่วนของตัวแปรเชิงปริมาณ ในกลุ่มของลักษณะที่เป็น 1

โดยที่ $p_1 = \frac{n_1}{n}$ เมื่อ $n = n_0 + n_1$ แทนจำนวนตัวแปรทั้งหมด

p_0 แทน ค่าสัดส่วนของตัวแปรเชิงปริมาณ ในกลุ่มของลักษณะที่เป็น 0

โดยที่ $p_0 = \frac{n_0}{n}$ เมื่อ $n = n_0 + n_1$ แทนจำนวนตัวแปรทั้งหมด

$\bar{s}_x$ แทน ส่วนเบี่ยงเบนมาตรฐานของตัวแปรเชิงปริมาณทั้งหมด (ทั้ง 2 กลุ่ม)

โดย $\bar{s}_x = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$ เมื่อ n แทนจำนวนตัวแปรทั้งหมด

3. เรียงลำดับยีนทั้ง p ยีน ตามค่าสัมบูรณ์ของ Point Biserial Correlation : $|r_{pb}|$ โดยเรียงจากค่า $|r_{pb}|$ ที่มากที่สุด ไปยังค่า $|r_{pb}|$ ที่น้อยที่สุด จะได้เซตของลำดับยีน (Rank List) ซึ่งแทนด้วยสัญลักษณ์ L นั่นคือ $L = \{x_1, x_2, \ldots, x_p\}$ โดยที่ $r_{pb}(x_j) = r_j$ เมื่อ $r_{pb}(x_j)$ แทน ค่าของ Point Biserial Correlation ของ x_j ซึ่งเขียนในรูปแบบอย่างง่ายคือ r_j

4. คำนวณหาค่า Enrichment Score (ES) ของแต่ละเซตของยีนจากข้อมูลที่มี : $ES^*(S_i)$

สำหรับ $i = 1, 2, , \ldots, t$

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

และ $$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(p - p_s)} \quad (2\text{-}3)$$

โดยที่ $N_R = \sum_{g_j \in S} |r_j|^\alpha$

α แทน การถ่วงน้ำหนักของยีนในเซตของยีน ซึ่งแสดงถึงความสัมพันธ์ของยีนกับฟีโนไทป์

p แทน จำนวนของยีนทั้งหมด

p_s แทน จำนวนของยีนในเซตของยีน

จะได้ $ES^*(S_i)$ = ค่าที่มากที่สุดที่ห่างจากศูนย์ของค่า $P_{hit} - P_{miss}$

$$ES^*(S_i) = \max\{ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i)\} \quad (2\text{-}4)$$

สำหรับ $i = 1, 2, , \ldots, t$

5. ทำการเรียงสับเปลี่ยน (permute) ในส่วนของค่าแสดงลักษณะของฟีโนไทป์ $\{0, 1\}$ ทั้งหมดแล้วกลับไปเริ่มทำในขั้นตอนที่ 2 ถึง ขั้นตอนที่ 5 ใหม่ โดยทำทั้งหมด m ครั้ง โดยที่

ครั้งที่ทำการเรียงสับเปลี่ยน			$ES^{(l)}(S_i)$
สำหรับ	1	$\Rightarrow$	$\{ES^{(1)}(S_i)\}$
สำหรับ	2	$\Rightarrow$	$\{ES^{(2)}(S_i)\}$
	$\vdots$		
สำหรับ	m	$\Rightarrow$	$\{ES^{(m)}(S_i)\}$

สำหรับทุก $i = 1, 2, , \ldots, t$ และ $l = 1, 2, , \ldots, m$

6. คำนวณหาค่า $p-value_i$ สำหรับแต่ละเซตของยีน (S_i) ทุก $i=1,2,,\ldots,t$ โดยที่

สำหรับ $ES^*(S_i)>0$ จะได้ว่า

$$p-value_i = \frac{\text{จำนวนของ } ES^{(l)}(S_i) \geq ES^*(S_i)}{m} \quad (2\text{-}5)$$

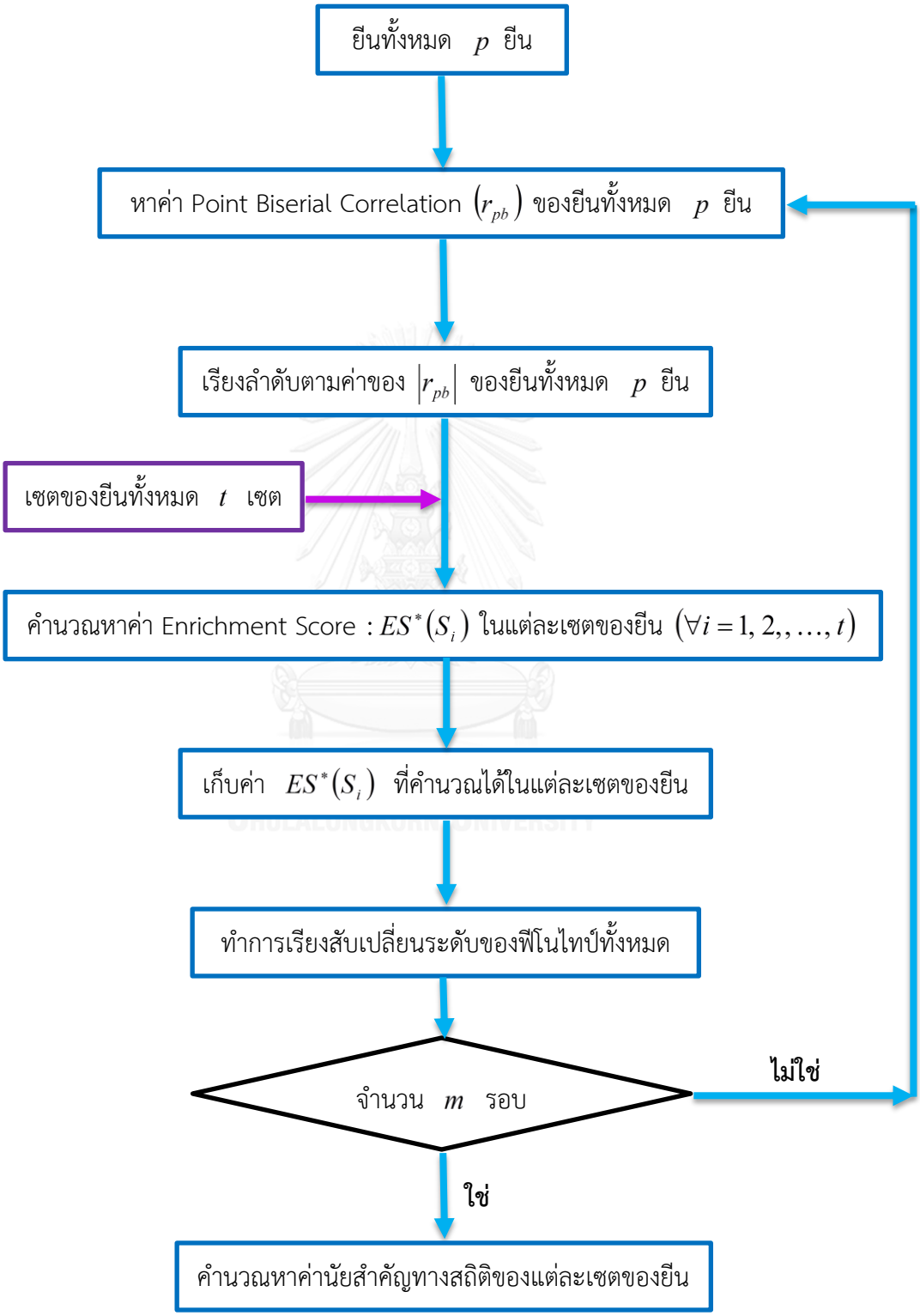
บางค่า $l=1,2,,\ldots,m$

สำหรับ $ES^*(S_i)<0$ จะได้ว่า

$$p-value_i = \frac{\text{จำนวนของ } ES^{(l)}(S_i) \leq ES^*(S_i)}{m} \quad (2\text{-}6)$$

บางค่า $l=1,2,,\ldots,m$

2.1.1 แผนภาพแสดงขั้นตอนของกระบวนการ GSEA



2.2 การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis)

การวิเคราะห์การถดถอยโลจิสติกเป็นการวิเคราะห์ที่มีวัตถุประสงค์ และแนวความคิดเหมือนกับการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression Analysis) กล่าวคือ เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ แต่มีความแตกต่างกันตรงที่ ตัวแปรตาม (Dependent Variable) ของการวิเคราะห์การการถดถอยเชิงเส้นตรงนั้นจะเป็นตัวแปรเชิงปริมาณ (Quantitative Variable) ในขณะที่ตัวแปรตามการของการวิเคราะห์การถดถอยโลจิสติกจะเป็นตัวแปรเชิงกลุ่ม (Categorical Variable) โดยตัวแปรตามที่เป็นตัวแปรเชิงกลุ่มนี้ อาจแบ่งออกได้เป็น 2 กลุ่ม (Dichotomous Data) หรือมากกว่า ซึ่งสำหรับการวิเคราะห์การถดถอยโลจิสติกกรณีที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่า การวิเคราะห์การถดถอยโลจิสติกทวิภาค (Binary Logistic Regression) (กัลยา วานิชย์บัญชา, 2552)

2.2.1 รูปแบบของการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

<u>*ตัวแปรตาม*</u> ; y เป็นตัวแปรเชิงกลุ่มที่แบ่งออกได้เป็น 2 กลุ่ม (Dichotomous Dependent Variable) กล่าวคือ $y_i \in \{0,1\}$ โดยที่ $y_i = \begin{cases} 1 & ; \text{เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ} \end{cases}$

เมื่อ n แทนจำนวนรายการของข้อมูล หรืออาจกล่าวได้ว่า $y_i \sim Bernoulli(\pi_i)\ ; i = 1,2,\ldots,n$
โดยที่ π_i คือ โอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และ $0 \le \pi_i \le 1$

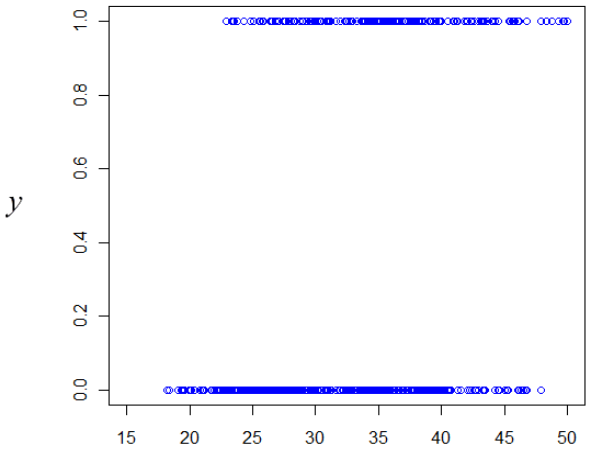
<u>*ตัวแปรอิสระ*</u> ; $x_1, x_2, \ldots, x_p$ เป็นตัวแปรเชิงปริมาณ หรือ ตัวแปรเชิงคุณภาพ
ดังนั้น p คือ จำนวนตัวแปรอิสระทั้งหมด

ในการวิเคราะห์การถดถอยโลจิสติกทวิภาคนั้น เมื่อพิจารณาถึงค่าคาดหวัง (Expected value) ของตัวแปรตาม (y) :

$$E[Y] = [0 \cdot P(y_i = 0)] + [1 \cdot P(y_i = 1)] = P(y_i = 1) = \pi_i \qquad (2\text{-}7)$$

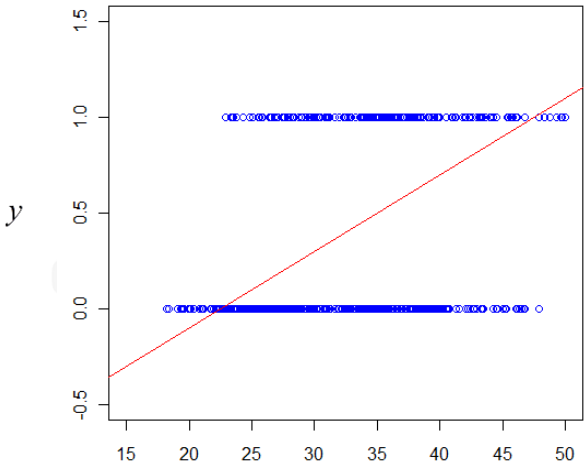
จะพบว่า ค่าคาดหวังของตัวแปรตาม y คือ ความน่าจะเป็น (Probability) ที่ตัวแปรตาม y มีค่าเท่ากับ 1 นั่นคือความน่าจะเป็นของเหตุการณ์หรือเรื่องที่เราสนใจนั่นเอง

เนื่องจากตัวแปรตาม y มีค่าที่เป็นไปได้แค่เพียง 2 ค่า คือ $\{0,1\}$ ดังตัวอย่างข้อมูล[2] ที่แสดงดังภาพที่ 2.2



ภาพที่ 2.2 แสดงค่าของข้อมูลตัวแปรตาม y ที่มีค่าเป็นเพียง 2 ค่า คือ 0 กับ 1

ซึ่งถ้าประมาณการลักษณะของข้อมูลข้างต้นด้วยกราฟเส้นตรง ดัง ภาพที่ 2.3 จะพบว่า การประมาณการด้วยกราฟเส้นตรงไม่เหมาะสมสำหรับข้อมูลในลักษณะนี้



ภาพที่ 2.3 แสดงการประมาณลักษณะของข้อมูลด้วยกราฟเส้นตรง

กล่าวคือ ถ้ากำหนดให้

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad ;i = 1, 2, \ldots, n \tag{2-8}$$

จะได้ว่า $\eta_i \in (-\infty, \infty) \quad ;i = 1, 2, \ldots, n$

---

[2] *Data : pima , From the faraway package of datasets in Extending the Linear Model with R book*

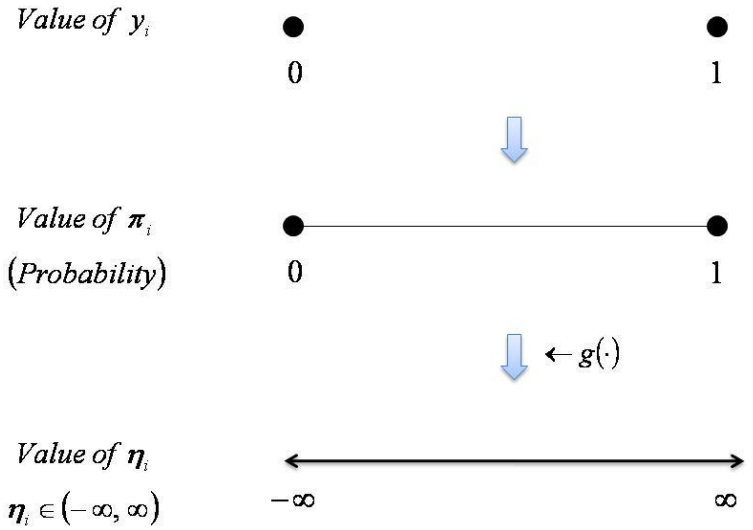
ซึ่งจากการที่ใช้สมการถดถอยเชิงเส้นตรง

$$y_i^{'} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad ;\forall i = 1, 2, \ldots, n \tag{2-9}$$

มาอธิบายหรือพยากรณ์ค่าของตัวแปรตาม y ใน ภาพที่ 2.2 นั้นจะเห็นได้ว่า ไม่สมเหตุสมผล เพราะว่า ค่าของตัวแปรตาม y มีค่าได้เพียงแค่ 2 ค่า ทำให้ค่าประมาณของ y เป็นโอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น นั่นก็คือค่าของ π_i ซึ่งมีค่าอยู่ในช่วง 0 ถึง 1 $(0 \le \pi_i \le 1)$ แต่ถ้าใช้สมการถดถอยเชิงเส้นตรงดังสมการที่ 2-9 ค่าของ y' ที่ได้อาจจะไม่ได้อยู่ในช่วง 0 ถึง 1 กล่าวคือ อาจมีค่าน้อยกว่า 0 หรือมากกว่า 1 ทั้งนี้เพราะว่าค่าของ y' มีค่าที่เป็นไปได้หลายค่าบนเส้นจำนวนจริง หรือกล่าวอีกนัยหนึ่งคือ กราฟของการวิเคราะห์การถดถอยโลจิสติกทวิภาคไม่ใช่กราฟเส้นตรงนั่นเอง ดังนั้น จำเป็นต้องหาฟังก์ชัน $g(\cdot)$ ที่เป็นตัวเชื่อมที่ทำให้

$$g(\pi_i) = \eta_i \tag{2-10}$$

ซึ่งสามารถแสดงให้เห็นได้ดังแผนภาพด้านล่าง



ภาพที่ 2.4 แสดงลักษณะของค่าที่เป็นไปได้ของตัวแปรตาม, ความน่าจะเป็น และ linear predictor

เราทราบว่า การแจกแจงแบบทวินาม (Binomial Distribution) สามารถจัดให้อยู่ในวงศ์ชี้กำลัง (Exponential Family) : $y_i \sim Binomial(n_i, \pi_i)$ จะได้ว่า

$$f(y_i|\theta) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \quad ;\forall i = 1, 2, \ldots, n \tag{2-11}$$

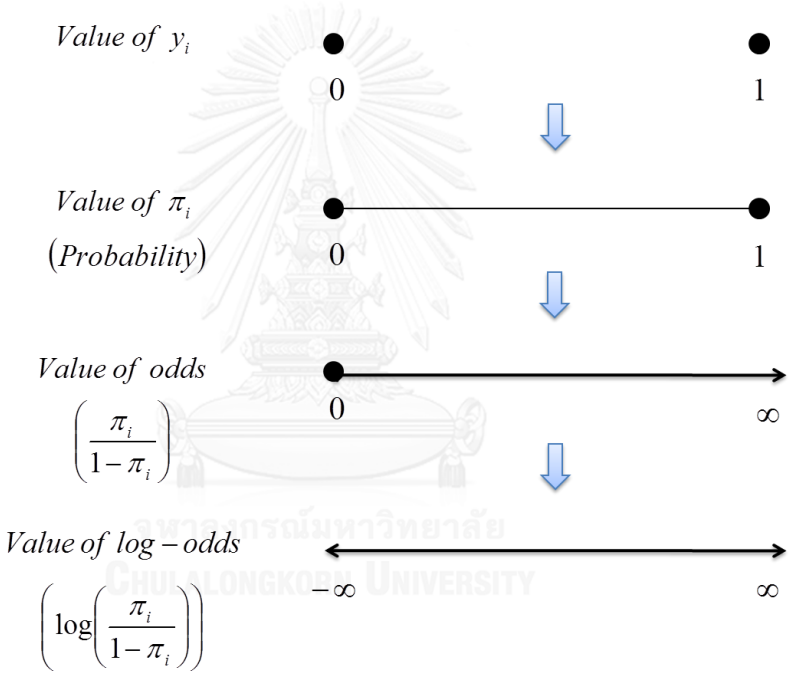
นั่นคือ สามารถเขียน

$$f(y_i|\theta)=\exp\left(y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+n_i\log(1-\pi_i)+\log\binom{n_i}{y_i}\right) \quad (2\text{-}12)$$

ซึ่งอยู่ในรูปของวงศ์ชี้กำลัง โดยที่ Canonical Link หรือ Link Function คือ Logit Link (Log-odds function) ซึ่งอยู่ในรูป

$$logit(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right) \quad (2\text{-}13)$$

โดยสามารถแสดงความสัมพันธ์ตามที่กล่าวมาได้ดังนี้



ภาพที่ 2.5 แสดงลักษณะของค่าที่เป็นไปได้ของตัวแปรตาม, ความน่าจะเป็น, odds และลอการึทึมของ odds

โดยกำหนดสมการดังสมการ 2-8 เมื่อ

$$g(\pi_i)=logit(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right) \quad (2\text{-}14)$$

ดังนั้น จะได้ว่าตัวแบบของการวิเคราะห์การถดถอยโลจิสติกทวิภาค คือ

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij} \quad ;\forall i=1,2,\ldots,n \quad (2\text{-}15)$$

โดยที่ ตัวผกผันของ Logit Function คือ Logistic Function ซึ่งอยู่ในรูป

$$\pi_i = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \qquad (2\text{-}16)$$

2.2.2 วิธีการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับการถดถอยโลจิสติกทวิภาค

ในการประมาณค่า (Estimate) ของพารามิเตอร์ (Parameters) $\beta_0, \beta_1, \ldots, \beta_p$ จะใช้วิธีการที่เรียกว่า *วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation Method : MLE)* โดยแนวความคิดของวิธี MLE มีอยู่ว่า การประมาณค่าพารามิเตอร์ (θ) ทำโดยอาศัยการสังเกตค่าที่วัดได้จากตัวอย่างสุ่มที่เลือกมาจากการแจกแจงที่ทราบรูปแบบของฟังก์ชันความหนาแน่น แต่ไม่ทราบค่าของพารามิเตอร์ (θ) จึงน่าจะสามารถใช้โอกาสที่เราจะเลือกตัวอย่าง และวัดค่าของหน่วยตัวอย่างต่างๆได้ มาพิจารณาค่าประมาณของพารามิเตอร์ θ ซึ่งโอกาสที่จะวัดค่าตัวอย่างสุ่มได้อาจจะแสดงได้ด้วยฟังก์ชันความหนาแน่นร่วมของค่าสังเกตของหน่วยตัวอย่าง $f(x;\theta)$ แต่ฟังก์ชันความหนาแน่นร่วมนี้ขึ้นอยู่กับค่าพารามิเตอร์ θ ดังนั้น ค่าประมาณของพารามิเตอร์ θ ที่น่าจะได้รับการพิจารณาก็คือ ค่าของพารามิเตอร์ θ ที่ทำให้ฟังก์ชันความหนาแน่นร่วมนี้มีค่าสูงสุด (สุชาดา กีระนันทน์, 2545)

เมื่อพิจารณาจากตัวแบบการถดถอยโลจิสติกทวิภาค ดังสมการที่ 2-12

กำหนดให้

$$\ell(\tilde{\beta}) = \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n}\left\{\binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}\right\} \qquad (2\text{-}17)$$

โดยที่ $\pi_i = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}$ และ $\tilde{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

จะได้ว่า *Log Likelihood Function* ของ $\ell(\tilde{\beta})$ คือ

$$\begin{aligned}\log \ell\left(\tilde{\beta}\right) &= \sum_{i=1}^{n} \log\left\{ \binom{n_i}{y_i} \pi_i^{y_i} \left(1-\pi_i\right)^{n_i - y_i} \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log\left(1-\pi_i\right) + \log\binom{n_i}{y_i} \right\}\end{aligned} \tag{2-18}$$

นั่นคือหาค่าประมาณของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ ที่ทำให้ $\log \ell\left(\tilde{\beta}\right)$ มีค่าสูงที่สุด ซึ่งจะสามารถหาได้โดยการหาอนุพันธ์ (Differentiate) ของฟังก์ชัน $\log \ell\left(\tilde{\beta}\right)$ เทียบกับพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ แล้วหาค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ ที่เป็นจุดวิกฤตของฟังก์ชันนี้ จะได้ว่า First order conditions คือ

$$\frac{\partial}{\partial \beta_k}\left(\log \ell\left(\tilde{\beta}\right)\right) = \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \right) \cdot x_{ik} = 0 \tag{2-19}$$

สำหรับทุก $k = 0, 1, \ldots, p$ โดยที่ $x_{i0} = 1$ ซึ่งจะมีสมการทั้งหมด $p+1$ สมการ

จะเห็นได้ว่า ในการหาค่าประมาณของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ โดยตรงด้วยวิธีนี้จะทำได้ค่อนข้างยาก ทั้งนี้เนื่องจากสมการทั้งหมด $p+1$ สมการนั้นเป็นสมการที่อยู่ในรูปที่ไม่เป็นเชิงเส้นตรง (Nonlinear Equation) ดังนั้นวิธีการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ อีกวิธีหนึ่งคือ *Newton-Raphson Method* ซึ่งเป็นกระบวนการทำซ้ำที่จะต้องกำหนดค่าของ β เริ่มต้น (initial value) เข้าไปในกระบวน จนกว่าขั้นตอนวิธีหรือคำสั่ง (algorithm) จะลู่เข้า (converge) สู่ค่าใดค่าหนึ่ง นั่นก็คือค่าประมาณของพารามิเตอร์ $\hat{\beta}$ นั่นเอง

2.2.3 การทดสอบสมมติฐาน

ในการวิเคราะห์การถดถอยโลจิสติกทวิภาค สามารถทำการทดสอบสมมติฐานได้ใน 2 ลักษณะ คือ การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธ์ความถดถอยโลจิสติกของตัวแปรอิสระแต่ละตัว และ การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธ์ความถดถอยโลจิสติกของตัวแปรอิสระหลายตัวพร้อมกัน หรืออาจกล่าวได้ว่าเป็นการทดสอบสมมติฐานระหว่างตัวแบบเต็มรูป (Full Model) กับตัวแบบลดรูป (Reduced Model) ซึ่งมีรายละเอียดดังต่อไปนี้

2.2.3.1 Wald Test

เป็นการทดสอบสมมติฐานทางสถิติของค่าสัมประสิทธิ์แต่ละตัวที่ได้จากการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ ว่ามีความแตกต่างจากศูนย์อย่างมีนัยสำคัญทางสถิติหรือไม่

โดยมีสมมติฐานคือ $H_0 : \beta_k = \beta_{k0}$ VS $H_a : \beta_k \neq \beta_{k0}$

ตัวสถิติสำหรับทดสอบสมมติฐานคือ

$$W = \frac{\hat{\beta}_k - \hat{\beta}_{k0}}{S\hat{E}\left(\hat{\beta}_k\right)} \sim N(0,1) \tag{2-20}$$

หรือ $$W^2 = \frac{\left(\hat{\beta}_k - \hat{\beta}_{k0}\right)^2}{\hat{Var}\left(\hat{\beta}_k\right)} \sim \chi_1^2 \tag{2-21}$$

2.2.3.2 Likelihood Ratio Test (LRT)

เป็นการทดสอบสมมติฐานทางสถิติของตัวแบบที่มีค่าสัมประสิทธิ์ที่ได้จากการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ เทียบกับตัวแบบที่ไม่มีค่าสัมประสิทธิ์ว่ามีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่

โดยมีสมมติฐานคือ $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ สำหรับ $1 \leq k \leq p$

VS H_a : มีอย่างน้อย 1 ค่าที่ $\beta_k \neq 0$; $k = 1, 2, \ldots, p$

ตัวสถิติสำหรับทดสอบสมมติฐานคือ

$$LRT = -2\log\left(\frac{\ell_0}{\ell_1}\right) = -2\log \ell_0 + 2\log \ell_1 \tag{2-22}$$

เมื่อ ℓ_0 คือ ค่าที่มากที่สุดของ Likelihood Function ภายใต้ H_0 หรือ Reduced Model

ℓ_1 คือ ค่าที่มากที่สุดของ Likelihood Function ภายใต้ H_a หรือ Full Model

โดยที่ $LRT \sim \chi_k^2$

2.3 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Penalized Regression

เป็นวิธีที่ค่อนข้างเป็นที่นิยมใช้กันอย่างแพร่หลายสำหรับการศึกษาข้อมูลที่มีมิติสูง (High-Dimensional Data) โดยในการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระ $\left(\tilde{\beta}\right)$ จะหาได้จากการหาค่า $\hat{\beta}$ ที่ทำให้ฟังก์ชันเป้าหมาย มีค่าต่ำสุด นั่นคือ

กำหนดฟังก์ชันเป้าหมาย

$$\left\| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right\|^2 + P_\lambda(\beta) \qquad ; i = 1, 2, \ldots, n \tag{2-23}$$

โดยที่

$$\hat{\beta} = \arg\min_{\beta} \left\| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right\|^2 + P_\lambda(\beta) \qquad ; i = 1, 2, \ldots, n \tag{2-24}$$

เมื่อ $P_\lambda(\beta)$ คือ Penalty Function

λ คือ Tuning Parameter ซึ่ง $\lambda \geq 0$

จะเห็นได้ว่าสมการข้างต้นนั้นมีความคล้ายคลึงกับวิธีกำลังสองน้อยที่สุดที่ใช้กันโดยทั่วไปสำหรับการหาค่า $\hat{\beta}$ แต่แตกต่างกันตรงที่ วิธี Penalized Regression จะมี Penalty Function เพิ่มขึ้นมาอีกพจน์หนึ่ง ซึ่งสำหรับค่าของ Tuning Parameter โดยทั่วไปแล้วจะใช้วิธี Cross-Validation ในการหาค่าที่เหมาะสมสำหรับข้อมูลที่ต้องการวิเคราะห์ ทั้งนี้หากเราเลือก Penalty Function ที่เหมาะสม ก็จะทำให้สมการข้างต้นนั้นสามารถคัดกรองตัวแปรเข้าในตัวแบบได้อีกด้วย กล่าวคือ Penalty Function นั้นจะทำให้สัมประสิทธิ์บางตัวในการประมาณค่ามีค่าเท่ากับศูนย์นั่นเอง

2.3.1 Penalty Function ของวิธี Least Absolute Shrinkage and Selection Operator (Lasso)

วิธี Lasso ได้ถูกเสนอโดย Tibshirani (1996) ซึ่งมีวัตถุประสงค์เพื่อใช้เป็นทั้งวิธีที่สามารถเลือกตัวแปรเข้าสู่ตัวแบบ และประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระ $\left(\tilde{\beta}\right)$ โดยวิธีนี้จะใช้ $\ell_1 - norm$ ในการปรับค่าด้วยวิธีกำลังสองน้อยที่สุด ซึ่งมี Penalty Function $\left(P_\lambda(\beta)\right)$ ดังนี้

$$P_{\lambda}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \tag{2-25}$$

ซึ่งค่า $\hat{\beta}$ หาได้จาก

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \left\| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad ; \forall i = 1, 2, \ldots, n \tag{2-26}$$

โดยตัวประมาณที่ได้จากวิธี Lasso จะมีค่า $\hat{\beta}$ ส่วนใหญ่เป็นศูนย์ และมีค่า $\hat{\beta}$ บางส่วนไม่เท่ากับศูนย์ (Sparse Estimator) ดังนั้น วิธี Lasso จึงเป็นวิธีที่สามารถเลือกตัวแปรเข้าได้โดยอัตโนมัติ (ตัวแปรที่ $\hat{\beta} \neq 0$) นั่นเอง นอกจากนี้แล้วการใช้วิธี Lasso ในการวิเคราะห์ข้อมูลที่มีมิติสูงยังมีข้อจำกัดอยู่ กล่าวคือ วิธี Lasso สามารถเลือกตัวแปรอิสระเข้าได้มากที่สุดจำนวน n ตัว ซึ่งถ้าข้อมูลในการวิเคราะห์ของเรามีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเป็นจำนวนมาก ทำให้ วิธี Lasso อาจจะไม่ค่อยเหมาะสมที่จะใช้ในการวิเคราะห์ และสำหรับกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง วิธี Lasso มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวเดียวจากกลุ่มของตัวแปรอิสระมีความสัมพันธ์กันสูงเข้าตัวแบบ โดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม โดย วิธี Lasso จะมีประสิทธิภาพสูงในการพยากรณ์ เมื่อตัวแบบจริงมีจำนวนของตัวแปรอิสระไม่มากนักที่มีความสัมพันธ์กับตัวแปรตาม และขนาดของ $\hat{\beta}$ ที่ไม่เท่ากับศูนย์มีขนาดใหญ่ (วิฐรา พึ่งพาพงศ์, 2558)

2.4 อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER)

เป็นโอกาสของการกระทำความผิดพลาดประเภทที่ 1 (Type I Error Rate : α) อย่างน้อยหนึ่งครั้งของชุดการเปรียบเทียบ หรือเป็นโอกาสของชุดการเปรียบเทียบ (set or family of contrasts) จำนวน 1 ชุด จะมีการตัดสินใจผิดพลาดประเภทที่ 1 เกิดขึ้น ซึ่งความผิดพลาดแบบนี้เกิดขึ้นในการเปรียบเทียบความแตกต่างค่าเฉลี่ยจำนวนหลายค่าหรือหลายกลุ่มค่าเฉลี่ย แล้วได้ข้อสรุปของการเปรียบเทียบดังกล่าวจำนวน 1 ชุด (set / family) (ไพฑูรย์ สุขศรีงาม 2557) โดยจะได้ว่าความผิดพลาดประเภทที่ 1 เกิดจากการที่ปฏิเสธสมมติฐานว่าง (Null Hypothesis : H_0) เมื่อสมมติฐานว่างเป็นจริง

ซึ่งอัตราความผิดพลาดรวมสามารถคำนวณได้จาก

$$FWER = P\left(\text{เกิด } Type\ I\ Error\right)$$

$$= \frac{\text{จำนวนของการจำลองที่เกิดความผิดพลาดประเภทที่ 1 อย่างน้อยหนึ่งครั้ง}}{\text{จำนวนของการจำลองทั้งหมด}} \tag{2-27}$$

(โดยในการทำการจำลองแต่ละครั้ง เราจะทำการปฏิเสธ H_0^C ก็ต่อเมื่อ $P_h^C < \alpha$ และค่าวัดประสิทธิภาพ FWER นี้ ยิ่งต่ำมากเท่าไร ก็จะยิ่งดี)

2.5 ค่าอำนาจในการทดสอบ (Power of Test)

ในกระบวนการขั้นตอนของการทดสอบสมมติฐานจะต้องตั้งสมมติฐานสองแบบ กล่าวคือ สมมติฐานว่าง (Null Hypothesis $: H_0$) ซึ่งโดยทั่วไปจะเป็นสมมติฐานที่ไม่มีการเปลี่ยนแปลงไปจากสภาวะเดิมหรือไม่มีผลที่แตกต่างจากของเดิม ส่วนอีกสมมติฐานหนึ่งคือ สมมติฐานทางเลือกอื่นหรือสมมติฐานแย้ง (Alternative Hypothesis $: H_a$) ซึ่งโดยทั่วไปจะเป็นสมมติฐานที่เกี่ยวกับความเชื่อที่ต้องการทดสอบ โดยในการสรุปผลมักจะเกิดความผิดพลาดได้สองแบบ กล่าวคือ

ความผิดพลาดประเภทที่ 1 (Type I Error) : α โดยที่ $\alpha = P\left(Reject\ H_0 | H_0\ is\ true\right)$ และ
ความผิดพลาดประเภทที่ 2 (Type II Error) : β โดยที่ $\beta = P\left(Accept\ H_0 | H_0\ is\ false\right)$

โดยทั่วไปแล้วปัญหาในการทดสอบสมมติฐานจะพยายามควบคุม α ให้มีค่าน้อย และจะพยายามทำให้ β มีค่าน้อยที่สุดเพื่อที่ทำให้ $1-\beta$ มีค่ามากที่สุด (ธีระพร วีระถาวร, 2536) ซึ่งเราเรียก $1-\beta$ ว่า อำนาจการทดสอบ โดยสำหรับการทดสอบใดๆที่มีค่าอำนาจการทดสอบยิ่งมาก จะแสดงว่าการทดสอบนั้นยิ่งดี ซึ่งค่าอำนาจการทดสอบสามารถคำนวณได้จาก

$$\text{POWER} = P\left(Reject\ H_0 | H_0\ is\ false\right)$$

$$= \frac{\text{จำนวนครั้งของการปฏิเสธ } H_0 \text{ เมื่อ } H_0 \text{ เป็นเท็จ}}{\text{จำนวนของ } H_0 \text{ ที่เป็นเท็จทั้งหมด}} \qquad (2\text{-}28)$$

บทที่ 3

วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA) และการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) ในการหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีน ซึ่งทำการศึกษาทั้งในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ และกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ หรือที่เรียกว่า “ข้อมูลที่มีมิติสูง (High-Dimensional Data)” โดยมีการจำลองข้อมูลของตัวแปรตามที่มีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution) และจำลองข้อมูลของตัวแปรอิสระแตกต่างกันใน 2 กรณี ได้แก่ ตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution) และ ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร(Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ และเมทริกซ์ความแปรปรวนร่วม รวมไปถึงกำหนดจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ใน 2 กรณีที่แตกต่างกันตามลักษณะความสัมพันธ์ของยีนในเซตยีน ได้แก่ กรณีที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และ กรณีที่มียีนแค่บางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา โดยทำการประมาณค่า p-value ของแต่ละเซตยีนในแต่ละวิธี ซึ่งในการเปรียบเทียบว่าวิธีการใดทำการประมาณค่า p-value ได้ดีและเหมาะสมกว่ากัน จะพิจารณาจาก 2 เกณฑ์ คือ อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test) โดยสำหรับการจำลองข้อมูลและการวิเคราะห์ข้อมูลทั้งหมดจะทำงานด้วยโปรแกรม R เวอร์ชัน 3.2.2 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาในส่วนของข้อมูลจำลองใน 2 กรณี ภายใต้ขอบเขตการวิจัยดังต่อไปนี้

3.1.1 ข้อมูลจำลอง : กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$

1. ทำการจำลอง (Simulate) ข้อมูลตัวแปรอิสระ

- ตัวแปรอิสระ (Independent Variables : $\tilde{x}_i$) : ยีน (genes) ทั้งหมด 30 ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด 100 ตัวอย่าง $(p=30, n=100)$ โดยกำหนดให้ข้อมูลตัวอย่างของแต่ละยีนทั้ง 30 ยีน มีการแจกแจงดังต่อไปนี้ :

  กรณีที่ 1 : ตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution)

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

  กรณีที่ 2 : ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

2. กำหนดกลุ่มของข้อมูลตามลักษณะความสัมพันธ์ของยีน $(Gene\ Set: S_i\ ; i=1,2,\ldots,6)$ ดังนี้

$$\left.\begin{aligned} Gene\ Set\ 1\ (S_1) &= \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_5\} \\ Gene\ Set\ 2\ (S_2) &= \{\tilde{x}_6, \tilde{x}_7, \ldots, \tilde{x}_{10}\} \\ Gene\ Set\ 3\ (S_3) &= \{\tilde{x}_{11}, \tilde{x}_{12}, \ldots, \tilde{x}_{15}\} \\ &\vdots \\ Gene\ Set\ 6\ (S_6) &= \{\tilde{x}_{26}, \tilde{x}_{27}, \ldots, \tilde{x}_{30}\} \end{aligned}\right\}$$

แต่ละเซตของยีนประกอบด้วย ยีนทั้งหมด 5 ยีน ในทุกๆเซตของยีน

3. ทำการกำหนดค่าสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระ โดยแบ่งเป็น 2 กรณีศึกษา ดังนี้

กรณีที่ 1 : ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ทั้งหมดโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระในกลุ่มของ

$Gene\ Set\ 1\ (S_1)$ 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

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

4. ทำการจำลอง (simulate) ข้อมูลตัวแปรตาม

- ตัวแปรตาม (Dependent Variables : y_i) : ลักษณะที่สนใจ (phenotype class)

ซึ่งแบ่งออกเป็น 2 ลักษณะกล่าวคือ $y_i = \begin{cases} 1 & ; \text{เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ} \end{cases}$

ซึ่ง $y_i \sim Bernoulli(\pi_i)\ ; i=1,2,\ldots,100$ เมื่อ $\pi_i = \dfrac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1+\exp(\sum_{j=1}^{p} \beta_j x_{ij})}$

โดยที่ π_i คือ โอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และ $0 \le \pi_i \le 1$

3.1.2 ข้อมูลจำลอง : กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

1. ทำการจำลอง (simulate) ข้อมูลตัวแปรอิสระ

- ตัวแปรอิสระ (Independent Variables : $\tilde{x}_i$) : ยีน (genes) ทั้งหมด 300 ยีน โดยแต่ละยีนประกอบไปด้วยข้อมูลตัวอย่าง (sample) ทั้งหมด 100 ตัวอย่าง $(p = 300, n = 100)$ โดยกำหนดให้ข้อมูลตัวอย่างของแต่ละยีนทั้ง 300 ยีน มีการแจกแจงดังต่อไปนี้ :

กรณีที่ 1 : ตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution)

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

กรณีที่ 2 : ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

2. กำหนดกลุ่มของข้อมูลตามลักษณะความสัมพันธ์ของยีน $(Gene\ Set: S_i\ ; i = 1,2,\ldots,60)$ ดังนี้

$$\begin{aligned} Gene\ Set\ 1\ (S_1) &= \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_5\} \\ Gene\ Set\ 2\ (S_2) &= \{\tilde{x}_6, \tilde{x}_7, \ldots, \tilde{x}_{10}\} \\ &\vdots \\ Gene\ Set\ 60\ (S_{60}) &= \{\tilde{x}_{296}, \tilde{x}_{297}, \ldots, \tilde{x}_{300}\} \end{aligned}$$

แต่ละเซตของยีนประกอบด้วย ยีนทั้งหมด 5 ยีน ในทุกๆเซตของยีน

3. ทำการกำหนดค่าสัมประสิทธิ์ $(\tilde{\beta})$ ของตัวแปรอิสระ โดยแบ่งเป็น 2 กรณีศึกษา ดังนี้

กรณีที่ 1 : ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ทั้งหมดโดยกำหนดค่าสัมประสิทธิ์ของตัวแปรอิสระมีค่าเป็น 1 สำหรับ 10 ตัวแปรอิสระในกลุ่มของ

$Gene\ Set\ 1\ (S_1)$ 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

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

4. ทำการจำลอง (simulate) ข้อมูลตัวแปรตาม

- ตัวแปรตาม (Dependent Variables : y_i) : ลักษณะที่สนใจ (phenotype class)

ซึ่งแบ่งออกเป็น 2 ลักษณะกล่าวคือ $y_i = \begin{cases} 1 & ; \text{เกิดเหตุการณ์ที่สนใจ} \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ} \end{cases}$

ซึ่ง $y_i \sim Bernoulli(\pi_i)\ ; i = 1,2,\ldots,100$ เมื่อ $\pi_i = \dfrac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij})}$

โดยที่ π_i คือ โอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และ $0 \le \pi_i \le 1$

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง
2. กำหนดค่าเริ่มต้น สำหรับการจำลองข้อมูล สำหรับแต่ละกรณีที่ทำการศึกษา
   2.1 กำหนดขนาดตัวอย่าง n
   2.2 กำหนดจำนวนตัวแปรอิสระ p ตัว
   2.3 กำหนดค่าสัมประสิทธิ์การถดถอยเริ่มต้น $(\tilde{\beta})$ สำหรับแต่ละกรณีที่ทำการศึกษา
   2.4 กำหนดระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระ (ρ)
3. จำลองข้อมูลจากค่าเริ่มต้นที่กำหนด
   3.1 จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติมาตรฐานหลายตัวแปร

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

   3.2 จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติหลายตัวแปร โดยมีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ $(\tilde{0})$ และเมทริกซ์ความแปรปรวนร่วม $\Sigma(p \times p)$

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

โดยที่

$$\Sigma = \begin{bmatrix} \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & 0 & \ldots & 0 \\ 0 & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \overbrace{\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}}^{5 \text{ ตัว}} \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$)

3.3 จำลองข้อมูลของตัวแปรตามที่มีการแจกแจงแบบแบร์นูลลี

$$y_i \sim Bernoulli(\pi_i); i = 1, 2, \ldots, n \quad \text{โดยที่} \quad \pi_i = \frac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij})}$$

4. นำข้อมูลที่ได้จากการจำลองมาทำการวิเคราะห์ตามขั้นตอนต่อไปนี้ สำหรับแต่ละกรณีที่ทำการศึกษา

4.1 กรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$

4.1.1 ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่า p-value ของแต่ละเซตของยีน

4.1.1.1 วิธี การถดถอยโลจิสติกทวิภาค

4.1.1.2 วิธี GSEA

4.2 กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

4.2.1 ใช้วิธี Lasso ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ (*เฉพาะสำหรับวิธีการถดถอยโลจิสติกทวิภาค*) โดยเลือกทั้งเซตของยีน ที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์ อย่างน้อย 1 ตัว

4.2.2 ใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคำนวณหาค่า p-value ของแต่ละเซตของยีน

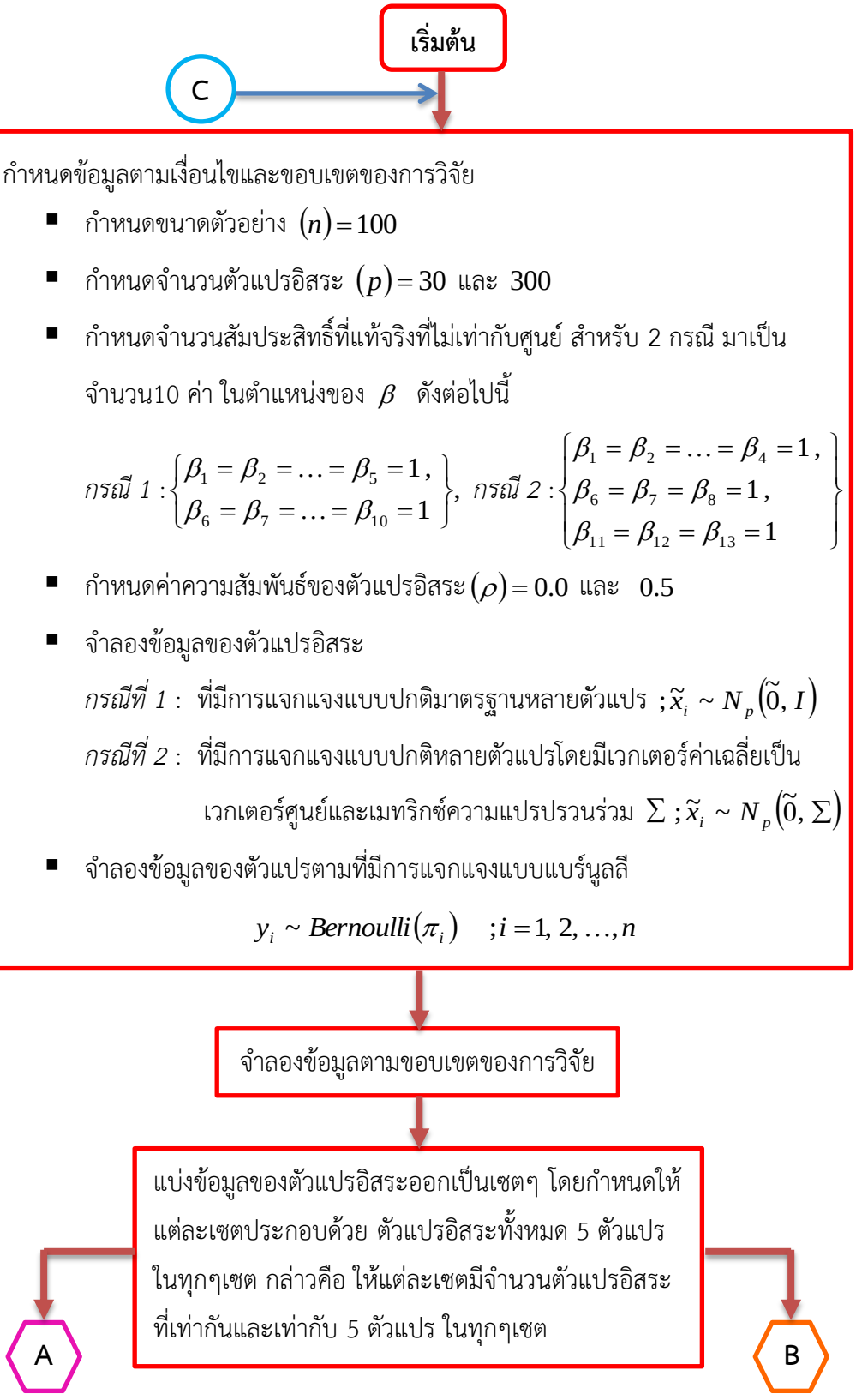
4.2.2.1 วิธีการถดถอยโลจิสติกทวิภาค (*สำหรับเซตของยีนที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระเท่ากับศูนย์ทุกตัว จะถือว่าค่า p-value ของเซตของยีนนั้นๆมีค่าเป็น 1*)

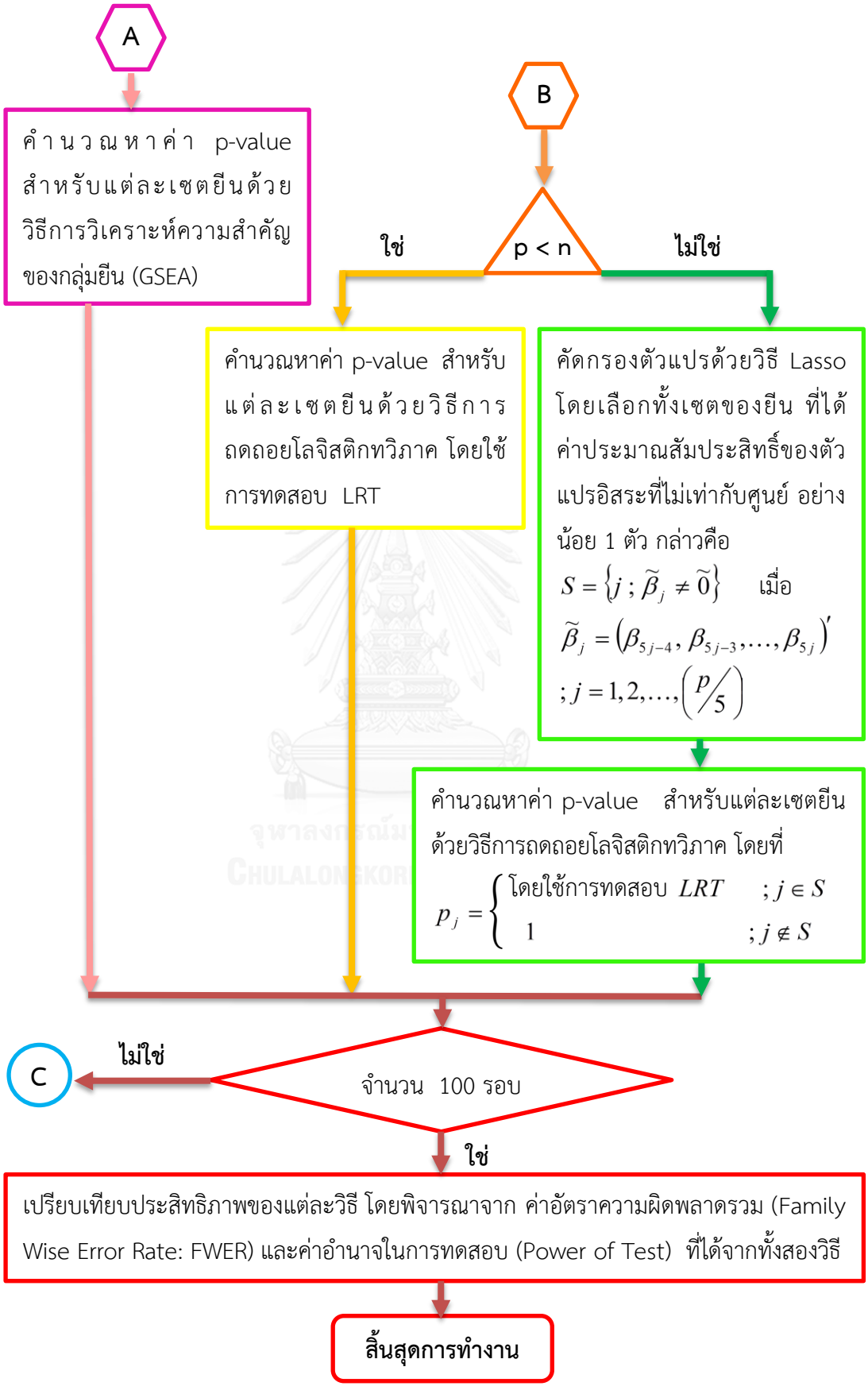
4.2.2.2 วิธี GSEA

5. นำข้อมูลที่ได้จากข้อ 4. มาคำนวณหาค่า อัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test)

6. วิเคราะห์ผลลัพธ์โดยทำการเปรียบเทียบค่าอัตราความผิดพลาดรวม FWER) และค่าอำนาจในการทดสอบ (Power of Test) ที่ได้จากทั้งสองวิธี โดยจำแนกตามลักษณะของข้อมูล (ขนาดตัวอย่าง และจำนวนของตัวแปรอิสระ), ลักษณะความสัมพันธ์ของยีนในเซตยีน และระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระที่ทำการศึกษา และสรุปผล

3.3 ขั้นตอนการทำงานของโปรแกรม




A
คำนวณหาค่า p-value สำหรับแต่ละเซตยีนด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (GSEA)
B
p < n
ใช่
ไม่ใช่
คำนวณหาค่า p-value สำหรับแต่ละเซตยีนด้วยวิธีการถดถอยโลจิสติกทวิภาค โดยใช้การทดสอบ LRT
คัดกรองตัวแปรด้วยวิธี Lasso โดยเลือกทั้งเซตของยีน ที่ได้ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์ อย่างน้อย 1 ตัว กล่าวคือ
$S = \{j ; \tilde{\beta}_j \neq \tilde{0}\}$ เมื่อ
$\tilde{\beta}_j = (\beta_{5j-4}, \beta_{5j-3}, \ldots, \beta_{5j})'$
$; j = 1, 2, \ldots, (p/5)$
คำนวณหาค่า p-value สำหรับแต่ละเซตยีนด้วยวิธีการถดถอยโลจิสติกทวิภาค โดยที่
$p_j = \begin{cases} \text{โดยใช้การทดสอบ } LRT & ; j \in S \\ 1 & ; j \notin S \end{cases}$
จำนวน 100 รอบ
ไม่ใช่
C
ใช่
เปรียบเทียบประสิทธิภาพของแต่ละวิธี โดยพิจารณาจาก ค่าอัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test) ที่ได้จากทั้งสองวิธี
สิ้นสุดการทำงาน

บทที่ 4

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของ 2 วิธีการสำหรับหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก ซึ่งได้แก่วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA) และการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) โดยการจำลองข้อมูลทั้งในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n=100,\ p=30)$ และกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ / ข้อมูลที่มีมิติสูง (High-Dimensional Data) $(n=100,\ p=300)$ ที่มีขอบเขตที่ต่างกัน ซึ่งจะพิจารณาในส่วนของความสัมพันธ์ (correlation) ของตัวแปรอิสระเป็น 0.0 และ 0.5 และลักษณะความสัมพันธ์ของยีนในเซตยีนซึ่งแบ่งเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าอัตราความผิดพลาดรวม (Family Wise Error Rate: FWER) และค่าอำนาจในการทดสอบ (Power of Test) โดยถ้าวิธีใดให้ค่าอัตราความผิดพลาดรวมต่ำที่สุด หรือมีค่าใกล้ 0 มากสุด และค่าอำนาจการทดสอบเฉลี่ยมากที่สุด จะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพและมีความเหมาะสมในการศึกษาหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักมากที่สุด

อักษรย่อและสัญลักษณ์ต่างๆที่ปรากฏในการนำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆแทนความหมายดังนี้

n	แทน ขนาดของตัวอย่าง
p	แทน จำนวนตัวแปรอิสระ
ρ	แทน ความสัมพันธ์ (correlation) ของตัวแปรอิสระ
β	แทน สัมประสิทธิ์การถดถอยของตัวแปรอิสระ
GSEA	แทน วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis)
Binary Logistic	แทน วิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

FWER	แทน ค่าอัตราความผิดพลาดรวม (Family Wise Error Rate)
POWER	แทน ค่าอำนาจในการทดสอบ (Power of Test)
Mean	แทน ค่าเฉลี่ย
S.D.	แทน ค่าเบี่ยงเบนมาตรฐาน

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบ โดยแบ่งออกเป็น 2 ส่วน คือ ในส่วนที่ 1 จะเปรียบเทียบค่าอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบสมมติฐานระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) สำหรับศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก และในส่วนที่ 2 จะเปรียบเทียบค่าอำนาจการทดสอบ (Power of Test) จากการทดสอบสมมติฐาน ระหว่าง 2 วิธีข้างต้น

โดยผลการวิจัยจะแบ่งออกเป็น 2 ส่วน ดังนี้

<u>ส่วนที่ 1</u> : ผลการเปรียบเทียบค่าอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบสมมติฐาน ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

เมื่อพิจารณาในกรณี

1. ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$
   - 1.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$
   - 1.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา
2. ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$
   - 2.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$

2.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

<u>ส่วนที่ 2</u> : ผลการเปรียบเทียบค่าอำนาจการทดสอบ (Power of Test) จากการทดสอบสมมติฐาน ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

เมื่อพิจารณาในกรณี

1. ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$

   1.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$

   1.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

2. ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

   2.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$

   2.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

4.1 ผลการเปรียบเทียบค่าอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบสมมติฐาน ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก จากวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) และเพื่อพิจารณาว่า

ปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาคในแต่ละวิธี ภายใต้ปัจจัย ดังต่อไปนี้

1. ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$
   1.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$
   1.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา
2. ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$
   2.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$
   2.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

โดยแสดงผลในตารางที่ 4.1.1 - 4.1.2 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ลักษณะของข้อมูล	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการศึกษาที่ต้องการเปรียบเทียบ
FWER	ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$	4.1.1	- ความสัมพันธ์ของตัวแปรอิสระ - ลักษณะของความสัมพันธ์ของยีนในเซตยีน	1. GSEA 2. Binary Logistic
	ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$	4.1.2		

ตารางที่ 4.1.1 แสดงค่าอัตราความผิดพลาดรวม (FWER) ของแต่ละสถานการณ์จากข้อมูลจำลองกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ 100 ชุด

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ค่าอัตราความผิดพลาดรวม			
	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$		$\rho = 0.5$	
	GSEA	Binary Logistic	GSEA	Binary Logistic
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	**0.01**	0.58	**0.00**	0.11
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	**0.00**	0.57	**0.00**	0.25

<u>หมายเหตุ</u> ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.1 ซึ่งแสดงผลของค่าอัตราความผิดพลาดรวม (FWER) โดยนับจากข้อมูลจำลองกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ 100 ข้อมูล ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) พบว่า

1) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.0

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น หาค่า FWER ได้เท่ากับ 0.01 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ดังนั้นวิธีการ

วิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น หาค่า FWER ได้เท่ากับ 0.00 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

2) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.5

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น หาค่า FWER ได้เท่ากับ 0.00 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค
- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น หาค่า FWER ได้เท่ากับ 0.00 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

ตารางที่ 4.1.2 แสดงค่าอัตราความผิดพลาดรวม (FWER) ของแต่ละสถานการณ์จากข้อมูลจำลอง กรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ 100 ชุด

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ค่าอัตราความผิดพลาดรวม			
	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$		$\rho = 0.5$	
	GSEA	Binary Logistic	GSEA	Binary Logistic
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	1.00	**0.08**	0.99	**0.00**
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	0.99	**0.08**	0.99	**0.01**

<u>หมายเหตุ</u> ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.2 ซึ่งแสดงผลของค่าอัตราความผิดพลาดรวม (FWER) โดยนับจากข้อมูลจำลองกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ 100 ข้อมูล ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) พบว่า

1) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.0

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก ด้วยวิธีการถดถอยโลจิสติกทวิภาคข้างต้น หาค่า FWER ได้เท่ากับ 0.08 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ดังนั้นวิธีการ

ถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการถดถอยโลจิสติกทวิภาคข้างต้น หาค่า FWER ได้เท่ากับ 0.08 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

2) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.5

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก ด้วยวิธีการถดถอยโลจิสติกทวิภาคข้างต้น หาค่า FWER ได้เท่ากับ 0.00 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการถดถอยโลจิสติกทวิภาคข้างต้น หาค่า FWER ได้เท่ากับ 0.01 ซึ่งต่ำกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

4.2 ผลการเปรียบเทียบค่าอำนาจการทดสอบ (Power of Test) จากการทดสอบสมมติฐานระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก จากวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาคในแต่ละวิธี ภายใต้ปัจจัย ดังต่อไปนี้

1. ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$
   1.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$
   1.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา
2. ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$
   2.1 เมื่อกำหนดความสัมพันธ์(correlation) ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.0$ และ $\rho = 0.5$
   2.2 เมื่อกำหนดลักษณะความสัมพันธ์ของยีนในเซตยีนเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

โดยแสดงผลในตารางที่ 4.2.1 - 4.2.2 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ลักษณะของข้อมูล	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการศึกษาที่ต้องการเปรียบเทียบ
POWER (เฉลี่ย)	ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$	4.2.1	- ความสัมพันธ์ของตัวแปรอิสระ - ลักษณะของความสัมพันธ์ของยีนในเซตยีน	1. GSEA 2. Binary Logistic
	ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$	4.2.2		

ตารางที่ 4.2.1 แสดงค่าเฉลี่ย (ค่าส่วนเบี่ยงเบนมาตรฐาน) ของค่าอำนาจการทดสอบ (POWER) ของแต่ละสถานการณ์จากข้อมูลจำลองกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ 100 ชุด

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ค่าอำนาจการทดสอบ (เฉลี่ย)			
	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$		$\rho = 0.5$	
	GSEA	Binary Logistic	GSEA	Binary Logistic
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	0.2950 (0.2472)	**0.9400 (0.2387)**	0.4750 (0.1095)	**1.0000 (0.0000)**
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	0.1900 (0.2024)	**0.9433 (0.1898)**	0.1567 (0.1738)	**0.9067 (0.1958)**

<u>หมายเหตุ</u> ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.1 ซึ่งแสดงผลของค่าอำนาจการทดสอบ (POWER) เฉลี่ย โดยนับจากข้อมูลจำลองกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ 100 ข้อมูล ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) พบว่า

1) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.0

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก ด้วยวิธีการถดถอยโลจิสติกทวิภาคข้างต้น สามารถหาค่า POWER เฉลี่ยได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ซึ่งเท่ากับ 0.9400 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าที่ค่อนข้างสูง ซึ่งเท่ากับ 0.2472 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนนั่นเอง ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค
- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการถดถอยโลจิสติกทวิภาคข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ซึ่งเท่ากับ 0.9433 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.2024 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนนั่นเอง ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

2) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.5

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วยวิธีการถดถอยโลจิสติกทวิภาคข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ซึ่งเท่ากับ 1.0000 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าที่ค่อนข้างสูง ซึ่งเท่ากับ 0.1095 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนนั่นเอง ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค
- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการถดถอยโลจิสติกทวิภาคข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน ซึ่งเท่ากับ 0.9067 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนมีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.1738 ถึงแม้ว่าจะต่ำกว่าของวิธีการถดถอยโลจิสติกทวิภาคก็ตาม แต่โดยเปรียบเทียบแล้วค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนถือว่ามีค่าที่สูงกว่า ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนนั่นเอง ดังนั้นวิธีการถดถอยโลจิสติกทวิภาคจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

ตารางที่ 4.2.2 แสดงค่าเฉลี่ย (ค่าส่วนเบี่ยงเบนมาตรฐาน) ของค่าอำนาจการทดสอบ (POWER) ของแต่ละสถานการณ์จากข้อมูลจำลองกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ 100 ชุด

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ค่าอำนาจการทดสอบ (เฉลี่ย)			
	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$		$\rho = 0.5$	
	GSEA	Binary Logistic	GSEA	Binary Logistic
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	**0.9200 (0.1842)**	0.1700 (0.3350)	**1.0000 (0.0000)**	0.0750 (0.2500)
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	**0.7467 (0.2375)**	0.0867 (0.2351)	**0.9300 (0.1365)**	0.0200 (0.1237)

<u>หมายเหตุ</u> ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.2 ซึ่งแสดงผลของค่าอำนาจการทดสอบ (POWER) เฉลี่ย โดยนับจากข้อมูลจำลองกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ 100 ข้อมูล ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) พบว่า

1) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.0

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ซึ่งเท่ากับ 0.9200

และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยโลจิสติกทวิภาค มีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.3350 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการถดถอยโลจิสติกทวิภาค นั่นเอง ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ซึ่งเท่ากับ 0.7467 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยโลจิสติกทวิภาค มีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.2351 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการถดถอยโลจิสติกทวิภาค นั่นเอง ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

2) ที่ระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระเท่ากับ 0.5

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วยวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ซึ่งเท่ากับ 1.0000 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยโลจิสติกทวิภาค มีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.2500 ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการถดถอยโลจิสติกทวิภาค นั่นเอง ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

- เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา การศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค โดยที่คำนึงถึงความสัมพันธ์ และการทำงานร่วมกันเป็นเซตของยีนเป็นหลักด้วย วิธีการวิเคราะห์ความสำคัญของกลุ่มยีนข้างต้น สามารถหาค่า POWER เฉลี่ย ได้สูงกว่าเมื่อเปรียบเทียบกับวิธีการถดถอยโลจิสติกทวิภาค ซึ่งเท่ากับ 0.9300 และเมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานจะเห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยโลจิสติกทวิภาค มีค่าที่ค่อนข้างสูง (สูงกว่าค่าเฉลี่ย) ซึ่งเท่ากับ 0.1237 ถึงแม้ว่าจะต่ำกว่าของวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนก็ตาม แต่โดยเปรียบเทียบแล้วค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการถดถอยโลจิสติกทวิภาคถือว่ามีค่าที่สูงกว่า ซึ่งแสดงถึงความผันผวนที่มากของค่า POWER เฉลี่ยของวิธีการถดถอยโลจิสติกทวิภาค นั่นเอง ดังนั้นวิธีการวิเคราะห์ความสำคัญของกลุ่มยีนจึงเป็นวิธีที่เหมาะสมกับการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบประสิทธิภาพของวิธีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาค ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) โดยจะพิจารณาในส่วนของความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 และ 0.5 และส่วนของลักษณะความสัมพันธ์ของยีนในเซตยีนซึ่งแบ่งเป็น 2 แบบ คือ ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา ซึ่งทำการศึกษาทั้งในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n=100,\ p=30)$ และกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ / ข้อมูลที่มีมิติสูง (High-Dimensional Data) $(n=100,\ p=300)$ โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าอัตราความผิดพลาดรวม (Family Wise Error Rate : FWER) และค่าอำนาจการทดสอบ (Power of Test) โดยสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 แบ่งผลการวิจัยออกเป็น 2 ส่วน โดยพิจารณาตามขนาดของตัวอย่าง ดังนี้

ส่วนที่ 1 : ผลการเปรียบเทียบอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบสมมติฐาน ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

ตารางที่ 5.1.1 แสดงวิธีการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาคที่เหมาะสมที่สุด เมื่อพิจารณาค่าอัตราความผิดพลาดรวม (FWER) ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 100 โดยจำแนกตามลักษณะของข้อมูล, ลักษณะความสัมพันธ์ของยีนในเซตยีน และระดับความสัมพันธ์ (correlation) ของตัวแปรอิสระที่ทำการศึกษา

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.0$	$\rho = 0.5$
	ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$		ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$	
	พิจารณาจากค่า FWER			
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	GSEA	GSEA	Binary Logistic	Binary Logistic
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	GSEA	GSEA	Binary Logistic	Binary Logistic

จากตารางที่ 5.1.1 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง (n) เท่ากับ 100 พิจารณาจากค่า FWER โดยค่าของ FWER จะได้วิธีที่เหมาะสมในการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาค ตรงกันข้ามกัน กล่าวคือ สำหรับกรณีที่ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ การศึกษาด้วยวิธี GSEA จะมีความเหมาะสม และสำหรับกรณีที่ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ การศึกษาด้วยวิธี Binary Logistic จะมีความเหมาะสม โดยสำหรับวิธี GSEA จะสามารถทำงานได้แย่ลง ในขณะที่วิธี Binary Logistic จะสามารถทำงานได้ดีขึ้น ในกรณีที่จำนวนของตัวแปรอิสระมีมากขึ้น ทั้งนี้ผลสรุปที่ได้เป็นผลภายใต้ขอบเขตการวิจัยที่ทำการศึกษา

ส่วนที่ 2 : ผลการเปรียบเทียบค่าอำนาจการทดสอบ (Power of Test) จากการทดสอบสมมติฐานระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)

ตารางที่ 5.1.2 แสดงวิธีการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาคที่เหมาะสมที่สุด เมื่อพิจารณาค่าอำนาจการทดสอบ (POWER) เฉลี่ย ระหว่างวิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis) และวิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 100 โดยจำแนกตามลักษณะของข้อมูล, ลักษณะความสัมพันธ์ของยีนในเซตยีน และระดับความสัมพันธ์(correlation) ของตัวแปรอิสระที่ทำการศึกษา

ลักษณะความสัมพันธ์ของยีนในเซตยีน	ความสัมพันธ์ (correlation) ของตัวแปรอิสระ			
	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.0$	$\rho = 0.5$
	ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$		ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$	
	พิจารณาจากค่า POWER (เฉลี่ย)			
ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด	Binary Logistic	Binary Logistic	GSEA	GSEA
มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา	Binary Logistic	Binary Logistic	GSEA	GSEA

จากตารางที่ 5.1.2 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง (n) เท่ากับ 100 พิจารณาจากค่า POWER เฉลี่ย โดยค่าของ POWER เฉลี่ยจะได้วิธีที่เหมาะสมในการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาค ตรงกันข้ามกัน กล่าวคือ สำหรับกรณีที่ขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ การศึกษาด้วยวิธี Binary Logistic จะมีความเหมาะสม และสำหรับกรณีที่ขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ การศึกษาด้วยวิธี GSEA จะมีความเหมาะสม โดยสำหรับวิธี GSEA จะสามารถทำงานได้ดีขึ้น ในขณะที่วิธี Binary Logistic จะสามารถทำงานได้แย่ลง ในกรณีที่จำนวนของตัวแปรอิสระมีมากขึ้น ทั้งนี้ผลสรุปที่ได้เป็นผลภายใต้ขอบเขตการวิจัยที่ทำการศึกษา

5.1.2 ผลจากความแตกต่างระหว่างลักษณะของข้อมูล (ขนาดตัวอย่าง และจำนวนของตัวแปรอิสระ)

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่า FWER สำหรับวิธี GSEA เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ จะดีกว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ ในขณะที่ประสิทธิภาพในการหาค่า FWER สำหรับวิธี Binary Logistic เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ จะแย่กว่าเมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ และประสิทธิภาพในการหาค่า POWER สำหรับวิธี GSEA เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ จะแย่กว่า เมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ ในขณะที่ประสิทธิภาพในการหาค่า POWER สำหรับวิธี Binary Logistic เมื่อขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ จะดีกว่า เมื่อขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$

5.1.3 ผลจากความแตกต่างระหว่างลักษณะความสัมพันธ์ของยีนในเซตยีน

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่า FWER สำหรับวิธี GSEA และ Binary Logistic เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีนที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด ไม่แตกต่างกันมากอย่างเห็นได้ชัดกับเมื่อลักษณะความสัมพันธ์ของยีนในเซตยีนที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา และประสิทธิภาพในการหาค่า POWER สำหรับวิธี GSEA และ Binary Logistic เมื่อลักษณะความสัมพันธ์ของยีนในเซตยีนที่ยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษาทั้งหมด จะค่อนข้างดีกว่าเมื่อลักษณะความสัมพันธ์ของยีนในเซตยีน ที่มียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา

5.1.4 ผลจากความแตกต่างของความสัมพันธ์ (correlation) ของตัวแปรอิสระ

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่า FWER สำหรับวิธี GSEA และ Binary Logistic เมื่อขนาดของความสัมพันธ์เป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างแย่กว่า เมื่อขนาดของความสัมพันธ์เป็น 0.5 และประสิทธิภาพในการหาค่า POWER สำหรับวิธี GSEA เมื่อขนาดของความสัมพันธ์เป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างแย่กว่าเมื่อขนาดของความสัมพันธ์เป็น 0.5 และสำหรับวิธี Binary Logistic เมื่อขนาดของความสัมพันธ์เป็น 0.0 โดยภาพรวมแล้วจะค่อนข้างดีกว่าเมื่อขนาดของความสัมพันธ์เป็น 0.5

5.2 สรุปผลโดยรวม

จากผลการวิจัยในการเปรียบเทียบวิธีการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาค ระหว่างวิธี GSEA และวิธี Binary Logistic สำหรับหาค่า p-value ของแต่ละเซตยีน โดยที่คำนึงถึงความสัมพันธ์ของทุกๆเซตของยีนเซต และการทำงานร่วมกันเป็นเซตของยีนเป็นหลัก ผลปรากฏว่าสำหรับกรณีขนาดตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ $(n > p)$ วิธี Binary Logistic มีอำนาจการทดสอบ (เฉลี่ย) สูง แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวม พบว่าวิธี GSEA มีค่าต่ำ ส่วนกรณีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ $(n < p)$ วิธี GSEA มีอำนาจการทดสอบ (เฉลี่ย) สูง แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวม พบว่าวิธี Binary Logistic มีค่าต่ำ ซึ่งจากทั้ง 2 กรณี เมื่อพิจารณาถึงทั้งสองค่าพร้อมกัน พบว่า วิธีการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาควิธีใดมีอำนาจการทดสอบ (เฉลี่ย) สูงก็จะมีอัตราความผิดพลาดรวมสูงด้วยเช่นกัน

ดังนั้นจึงไม่สามารถสรุปได้ว่าวิธีการในการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาควิธีใดเป็นวิธีที่ดีที่สุด อย่างไรก็ตาม สำหรับการนำไปใช้งานจริง จะขึ้นอยู่กับผู้นำไปใช้ว่าจะพิจารณาให้ความสำคัญกับประสิทธิภาพของวิธีศึกษาด้วยค่าอัตราความผิดพลาดรวม หรือ ค่าอำนาจการทดสอบ มากกว่ากัน แล้วถึงเลือกใช้วิธีการศึกษาความสัมพันธ์ของเซตของยีนและฟีโนไทป์แบบทวิภาคที่เหมาะสม ทั้งนี้ผลสรุปที่ได้เป็นผลภายใต้ขอบเขตการวิจัยที่ทำการศึกษา

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ สำหรับผู้ที่สนใจอาจจะนำไปศึกษาต่อได้อีกในเรื่องของ

1. การกำหนดความสัมพันธ์ให้แต่ละเซตของยีนอาจกำหนดให้มีจำนวนยีนที่แตกต่างกัน ซึ่งในความเป็นจริงแล้วยีนเซตแต่ละเซตอาจมีจำนวนยีนอยู่แตกต่างกันได้
2. ขอบเขตในการวิจัย ในเรื่องของลักษณะของข้อมูล (ขนาดตัวอย่าง และจำนวนของตัวแปรอิสระ), ลักษณะความสัมพันธ์ของยีนในเซตยีน และความสัมพันธ์ (correlation) ของตัวแปรอิสระ อาจจะมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้นได้
3. กรณีที่ตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่แบ่งออกได้มากกว่า 2 กลุ่ม
4. กรณีที่รูปแบบความสัมพันธ์ของตัวแปรอิสระเปลี่ยนไป
5. จำนวนรอบของการรันโปรแกรม 100 รอบอาจไม่ใช่จำนวนรอบที่ดีที่สุด ดังนั้นอาจมีการเพิ่มจำนวนรอบให้มากยิ่งขึ้นได้อีก
6. วิธีการคัดกรองตัวแปรในความเป็นจริงแล้วยังมีอีกหลายวิธีที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการคัดกรองตัวแปรอื่นๆมาใช้ในการพิจารณาได้อีก
7. เกณฑ์ที่ใช้ในการตัดสินใจสำหรับเปรียบเทียบประสิทธิภาพของแต่ละวิธี อาจต้องมีการใช้เกณฑ์ในการตัดสินใจเพิ่มมากกว่านี้เพื่อที่จะสามารถสรุปผลให้ชัดเจนได้ว่าวิธีใดที่มีประสิทธิภาพมากที่สุดในสองวิธี

รายการอ้างอิง

ภาษาไทย

กัลยา วานิชย์บัญชา. (2552). การวิเคราะห์ข้อมูลหลายตัวแปร: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

ธีระพร วีระถาวร. (2536). การอนุมานเชิงสถิติขั้นกลาง โครงสร้างและความหมาย. กรุงเทพฯ : ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

ไพฑูรย์ สุขศรีงาม (2557). วิธีการเปรียบเทียบความแตกต่างค่าเฉลี่ยรายคู่หลังการวิเคราะห์ความแปรปรวน. ว.มรม.(มนุษยศาสตร์และสังคมศาสตร์) ปีที่ 8, 1, 23 – 30.

วิฐรา พึ่งพาพงศ์. (2558). บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี ปีที่ 23, 2, 212 – 223.

สุชาดา กีระนันทน์. (2545). การอนุมานเชิงสถิติ : ทฤษฎีขั้นต้น: กรุงเทพฯ : โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Benjamini, Y. a. Y. H. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society,* 57, 289-300.

Howell, B. S. E. D. C. (2005). Point Biserial Correlation *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1552-1553).

Subramanian, A., Tamayo, P., K.Mootha, V., Mukherjee, S., L.Ebert, B., A.Gillette, M., . . . P.Mesirov, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America (PNAS),* 102, 15545 -15550.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society,* 58, 267-288.

ภาคผนวก ก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่าง <u>กรณีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ</u> โดยมีขนาดตัวอย่างเท่ากับ 100, จำนวนตัวแปรอิสระเท่ากับ 30 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 และยีนทุกตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ทั้งหมด โดยมีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาคด้วยวิธี

- วิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)
- วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA)

```
library(mvtnorm)
library(ltm)
library(glmnet)
library(Matrix)

n<-100
p<-30
rho<-0
S<-p/5

########## Mean and Sigma for X ##########

mean<-matrix(c(numeric(p)), nrow=p, ncol=1)
Sigma<-diag(p)
for(c in 1:S)
  {
    Sigma[(c*5-4):(c*5), (c*5-4):(c*5)]<-rho
  }
```

```
diag(Sigma)<-1

########### Group Gene Set ################

GS<-function(e=seq(1:S))
                {
                   G<-matrix( )
                   G<-X[1:100, (e*5-4):(e*5)]
                   G<-as.matrix(G)
                   return(G)
                 }
#########################################################

iterate<-100

P_value.GSEA<-matrix(0, iterate, S)
Type1Error.GSEA<-matrix(0, iterate, 1)
Power.GSEA<-matrix(0, iterate, 1)

P_value.Logistic<-matrix(0, iterate, S)
Type1Error.Logistic<-matrix(0, iterate, 1)
Power.Logistic<-matrix(0, iterate, 1)

aa<-1
repeat
{
```

```
############ Simulation Data X ############

X<-rmvnorm(n, mean, Sigma)

############ Simulation Beta ##############

beta<-matrix(0, p, 1)
fixbeta<-c(1:5, 6:10)
beta[fixbeta, ]<-1
B<-as.matrix(beta)

############ Simulation Data Y ############

eta<-X%*%B
prob<-exp(eta)/(1+exp(eta))
Y<-rbinom(n, size=1, prob=prob)
Y<-as.matrix(Y)

data<-data.frame(Y=Y, X=X)

##### index in each gene set #####

index<-1:p
set<-matrix(0, p, 1)
for(f in 1:S)
   {
      set[(f*5-4):(f*5), ]<-f
   }
```

```
geneset<-data.frame(index, set)

############################################################

############ Logistic ############

Fmodel<-glm(Y~X, family=binomial, data=data)
DF <- deviance(Fmodel)
dfF<-df.residual(Fmodel)

pvalue.Logistic<-matrix(0, S, 1)
for(g in 1:S)
  {
     X.R<-X[ , -((g*5-4):(g*5))]
     Rmodel<-glm(Y~X.R, family=binomial, data=data)
     DR <-deviance(Rmodel)
     dfR<-df.residual(Rmodel)
     LRT<-DR-DF
     pvalue.Logistic[g]<-pchisq(LRT, dfR-dfF, lower.tail=F)
     pvalue.Logistic<-as.matrix(pvalue.Logistic)
  }
P_value.Logistic[aa, ]<-pvalue.Logistic[ ,1]

############################################################
```

```
############ GSEA ############
##### Gene Expression Data ######

DE_0<-data[-which(data$Y=="1"), ]
DE_1<-data[-which(data$Y=="0"), ]
Class0<-t(DE_0)
Class1<-t(DE_1)
DiffExpressM<-cbind(Class0, Class1)

##### Point Biserial Correlation #####

r.pb<-rep(0, p)
for(h in 1:p)
   {
      r.pb[h]<-abs(biserial.cor(X[ ,h], as.vector(Y)))
   }

##### Rank List #####

Names<-colnames(data)[2:31]
genediff<-data.frame(Names, r.pb)
Ranklist<-genediff[with(genediff, order(-r.pb, Names)), ]
order_r<-order(r.pb, decreasing=T)

phit<-matrix(rep(NA, S*p), nrow=S)
pmiss<-matrix(rep(NA, S*p), nrow=S)

for(s in 1:S)
```

```
  {
     for(i in 1:p)
        {
          # calculate phit
          hitindex<-intersect(geneset$index[geneset$set==s], order_r[1:i])
      phit[s,i]<-sum(abs(r.pb[hitindex]))/sum(abs(r.pb[geneset$index[geneset$set==s]]))

          # calculate pmiss
          missindex<-intersect(geneset$index[geneset$set!=s], order_r[1:i])
          pmiss[s,i]<-0
                for(j in 1:length(missindex))
                   {
              pmiss[s,i]<-pmiss[s,i]+
                   (1/(p - length(geneset$set[geneset$set==geneset$set[missindex[j]]])))
                   }
        }
  }

#### ES1-ES6(observed) ######

ES.obs<-matrix(0, S, 1)
for(k in 1:S)
   {
     d.obs<-phit-pmiss
     ES.obs[k]<-if(max(d.obs[k, ])==max(abs(d.obs[k, ]))){max(d.obs[k, ])}
                  else {min(d.obs[k, ])}
   }
```

```
#### Permute Y >>> 100 times #####

M<-100
YY<-sample(Y[ ,1], 100)
ES<-numeric( )
for(m in 1:M)
    {
        YY<-sample(Y[ ,1], 100)
        r.pb<-rep(0, p)
        for(h in 1:p)
           {
              r.pb[h]<-abs(biserial.cor(X[ ,h], as.vector(YY)))
           }
        Names<-colnames(data)[2:31]
        genediff<-data.frame(Names, r.pb)
        RL<-genediff[with(genediff,order(-r.pb, Names)),]

        index<-1:p
        set<-matrix(0, p, 1)
        for(f in 1:S)
           {
              set[(f*5-4):(f*5), ]<-f
           }
        geneset<-data.frame(index, set)
        order_r<-order(r.pb, decreasing=T)

        phit<-matrix(rep(NA, S*p), nrow=S)
        pmiss<-matrix(rep(NA, S*p), nrow=S)
```

```
    for (s in 1:S)
      {
        for (i in 1:p)
            {
             # calculate phit
          hitindex<-intersect(geneset$index[geneset$set==s], order_r[1:i])
          phit[s,i]<-sum(abs(r.pb[hitindex]))/
                    sum(abs(r.pb[geneset$index[geneset$set==s]]))

           # calculate pmiss
          missindex<-intersect(geneset$index[geneset$set!=s], order_r[1:i])
          pmiss[s,i]<-0
          for (j in 1:length(missindex))
              {
                  pmiss[s,i]<-pmiss[s,i]+
             (1/(p- length(geneset$set[geneset$set==geneset$set[missindex[j]]])))
              }
           }
       }
   es<-matrix(0,S,1)
   for(k in 1:S)
      {
        d<-phit-pmiss
        es[k]<-if(max(d[k, ])== max(abs(d[k, ])))max(d[k, ])} else {min(d[k, ])}
      }
   ES<-rbind(ES, es)
  }
ESS<-matrix(rep(NA, S*M), nrow=S)
```

```
for (v in 1:S)
     {
      for (w in 1:M)
           {
               ESS[v,w]<-ES[S*w-S+v, ]
           }
    }

#### compute P-Value ####

pvalue.GSEA<-matrix(0, S, 1)
for(t in 1:S)
   {
      pvalue.GSEA[t]<-if(ES.obs[t, ]>0) {mean(ESS[t, ]>=ES.obs[t, ])}
                          else {mean(ESS[t, ]<=ES.obs[t, ])}
      pvalue.GSEA<-as.matrix(pvalue.GSEA)
   }
P_value.GSEA[aa, ]<-pvalue.GSEA[ ,1]

############################################################
############## Type I Error ###################

type1.e.G<-0
type1.e.L<-0

   tmp1<-sum(pvalue.GSEA[3:6]<0.05)
        { if(tmp1>0)  type1.e.G<-type1.e.G+1 }
```

```
    tmp2<-sum(pvalue.Logistic[3:6]<0.05)
        { if(tmp2>0) type1.e.L<-type1.e.L+1 }

Type1Error.GSEA[aa, ]<-type1.e.G
Type1Error.Logistic[aa, ]<-type1.e.L

############## Power of Test ###################

      pow.G<-length(pvalue.GSEA[1:2][pvalue.GSEA[1:2]<0.05])
      pow.L<-length(pvalue.Logistic[1:2][pvalue.Logistic[1:2]<0.05])

Power.GSEA[aa, ]<-pow.G/2
Power.Logistic[aa, ]<-pow.L/2

aa <- aa + 1
if(aa > iterate) {break}
}

P_value.GSEA<-data.frame(P_value.GSEA=P_value.GSEA)
write.table(P_value.GSEA, file = "P_value.GSEAc1_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

P_value.Logistic<-data.frame(P_value.Logistic=P_value.Logistic)
write.table(P_value.Logistic, file = "P_value.Logisticc1_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

TypeI.Error<-data.frame(Type1Error.GSEA= Type1Error.GSEA, Type1Error.Logistic=
Type1Error.Logistic)
```

```
write.table(TypeI.Error, file = "Type1Errorc1_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

FWER.GSEA <- nnzero(as.numeric(Type1Error.GSEA), na.counted = FALSE)/iterate
FWER.Logistic <- nnzero(as.numeric(Type1Error.Logistic), na.counted = FALSE)/iterate
FWER<-data.frame(FWER.GSEA=FWER.GSEA, FWER.Logistic=FWER.Logistic)
write.table(FWER, file = "FWERc1_final.csv", sep = ",", col.names = TRUE,qmethod =
"double")

power<-data.frame(Power.GSEA= Power.GSEA, Power.Logistic= Power.Logistic)
ave.power.GSEA<-mean(power[ ,1])
sd.power.GSEA<-sd(power[ ,1])
ave.power.Logistic<-mean(power[ ,2])
sd.power.Logistic<-sd(power[ ,2])
stats.power<-data.frame(ave.power.GSEA, sd.power.GSEA, ave.power.Logistic,
sd.power.Logistic)

write.table(power, file = "powerc1_final.csv", sep = ",", col.names = TRUE,qmethod =
"double")
write.table(stats.power, file = "statspowerc1_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")
```

ภาคผนวก ข

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่าง กรณีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ โดยมีขนาดตัวอย่างเท่ากับ 100, จำนวนตัวแปรอิสระเท่ากับ 300 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0.5 และมียีนบางตัวในกลุ่มมีความสัมพันธ์กับฟีโนไทป์ที่ต้องการศึกษา โดยมีการศึกษาความสัมพันธ์ระหว่างเซตของยีนและฟีโนไทป์แบบทวิภาคด้วยวิธี

- วิธีการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis)
- วิธีการวิเคราะห์ความสำคัญของกลุ่มยีน (Gene Set Enrichment Analysis : GSEA)

```
library(mvtnorm)
library(ltm)
library(glmnet)
library(Matrix)

n<-100
p<-300
rho<-0.5
S<-p/5

########## Mean and Sigma for X ##########

mean<-matrix(c(numeric(p)), nrow=p, ncol=1)

Sigma<-diag(p)
for(c in 1:S)
   {
     Sigma[(c*5-4):(c*5), (c*5-4):(c*5)]<-rho
   }
diag(Sigma)<-1
```

```
############ Group Gene Set ################

GS<-function(e=seq(1:S))
                {
                  G<-matrix( )
                  G<-X[1:100, (e*5-4):(e*5)]
                  G<-as.matrix(G)
                  return(G)
                }

#########################################################

iterate<-100

P_value.GSEA<-matrix(0, iterate, S)
Type1Error.GSEA<-matrix(0, iterate, 1)
Power.GSEA<-matrix(0, iterate, 1)

P_value.Logistic<-matrix(0, iterate, S)
Type1Error.Logistic<-matrix(0, iterate, 1)
Power.Logistic<-matrix(0, iterate, 1)

aa<-1
repeat
{

############ Simulation Data X ############

X<-rmvnorm(n, mean, Sigma)
```

```
############ Simulation Beta #############

beta<-matrix(0, p, 1)
fixbeta<-c(1:4, 6:8, 11:13)
beta[fixbeta, ]<-1
B<-as.matrix(beta)

############ Simulation Data Y ############

eta<-X%*%B
prob<-exp(eta)/(1+exp(eta))
Y<-rbinom(n, size=1, prob=prob)
Y<-as.matrix(Y)

data<-data.frame(Y=Y, X=X)

##### index in each gene set #####

index<-1:p
set<-matrix(0, p, 1)
for(f in 1:S)
   {
     set[(f*5-4):(f*5), ]<-f
   }
geneset<-data.frame(index, set)

############################################################
```

```
############ Logistic ############

z<-dim(X)[2]
cv<-cv.glmnet(x=X, y=Y, nfolds=10, family="binomial", standardize=FALSE)
model<-glmnet(x=X, y=Y, family="binomial", lambda=cv$lambda.min,
standardize=FALSE)
b<-coef(model)
b<-b[-1, 1]
beta.lasso<-rep(0, z)
beta.lasso[which(b!=0)]<-b[which(b!=0)]
b.lasso.set<-data.frame(beta.lasso, set)
b.lasso.sig<-subset(b.lasso.set, beta.lasso!=0, select=set)

if(nrow(b.lasso.sig)==0) { aa<- aa + 0}
else {
        sigset<-unique(b.lasso.sig)

        Xlasso<-matrix(0, n, (nrow(sigset)*5))
        for(u in 1:nrow(sigset))
            {
               q<-sigset[u,1]
               Xlasso[ ,(u*5-4):(u*5)]<-GS(q)
               Xlasso<-as.matrix(Xlasso)
            }
        datanew<-data.frame(Y=Y, X=Xlasso)

        Fmodel<-glm(Y~Xlasso, family=binomial, data=datanew)
        DF <- deviance(Fmodel)
        dfF<-df.residual(Fmodel)

        pvalue.Logistic<-matrix(1, S, 1)
```

```
if(nrow(sigset)==1) {
      for(r in 1:nrow(sigset))
         {
            q<-sigset[r,1]
            Rmodel<-glm(Y~1, family=binomial)
            DR <-deviance(Rmodel)
            dfR<-df.residual(Rmodel)
            LRT<-DR-DF
            pvalue.Logistic[q]<-pchisq(LRT, dfR-dfF, lower.tail=F)
            pvalue.Logistic<-as.matrix(pvalue.Logistic)
         }
                     }
else if(nrow(sigset)>1 && nrow(sigset)<=(n/5))
     {
        for(r in 1:nrow(sigset))
            {
               q<-sigset[r,1]
               Xlasso.R<-Xlasso[ ,-((r*5-4):(r*5))]
               Rmodel<-glm(Y~Xlasso.R, family=binomial, data=datanew)
               DR <-deviance(Rmodel)
               dfR<-df.residual(Rmodel)
               LRT<-DR-DF
               pvalue.Logistic[q]<-pchisq(LRT, dfR-dfF, lower.tail=F)
               pvalue.Logistic<-as.matrix(pvalue.Logistic)
            }
     }
     else { aa <- aa + 0 }

P_value.Logistic[aa, ]<-pvalue.Logistic[ ,1]

############################################################
```

```
############ GSEA ############

##### Gene Expression Data ######
        DE_0<-data[-which(data$Y=="1"), ]
        DE_1<-data[-which(data$Y=="0"), ]
        Class0<-t(DE_0)
        Class1<-t(DE_1)
        DiffExpressM<-cbind(Class0, Class1)

##### Point Biserial Correlation #####

        r.pb<-rep(0, p)
        for(h in 1:p)
           {
              r.pb[h]<-abs(biserial.cor(X[ ,h], as.vector(Y)))
           }

##### Rank List #####

        Names<-colnames(data)[2:301]
        genediff<-data.frame(Names, r.pb)
        Ranklist<-genediff[with(genediff,order(-r.pb, Names)),]
        order_r<-order(r.pb, decreasing=T)

        phit<-matrix(rep(NA, S*p), nrow=S)
        pmiss<-matrix(rep(NA, S*p), nrow=S)
        for(s in 1:S)
           {
             for(i in 1:p)
                 {
                   # calculate phit
```

```
                hitindex<-intersect(geneset$index[geneset$set==s], order_r[1:i])
                phit[s,i]<-sum(abs(r.pb[hitindex]))/
                            sum(abs(r.pb[geneset$index[geneset$set==s]]))

                # calculate pmiss
                missindex<-intersect(geneset$index[geneset$set!=s], order_r[1:i])
                pmiss[s,i]<-0
                for(j in 1:length(missindex))
                    {
                      pmiss[s,i]<-pmiss[s,i]+
                    (1/(p -length(geneset$set[geneset$set==geneset$set[missindex[j]]])))
                     }
                }
            }

#### ES1 - ES60 (observed) ######

       ES.obs<-matrix(0, S, 1)
       for(k in 1:S)
           {
              d.obs<-phit-pmiss
              ES.obs[k]<-if(max(d.obs[k, ])== max(abs(d.obs[k, ]))) {max(d.obs[k, ])}
                          else {min(d.obs[k, ])}
           }

#### Permute Y >>> 100 times #####

        M<-100
        YY<-sample(Y[ ,1], 100)
        ES<-numeric( )
```

```
for(m in 1:M)
  {
     YY<-sample(Y[ ,1], 100)
     r.pb<-rep(0, p)
     for(h in 1:p)
        {
           r.pb[h]<-abs(biserial.cor(X[ ,h], as.vector(YY)))
        }
     Names<-colnames(data)[2:301]
     genediff<-data.frame(Names, r.pb)
     RL<-genediff[with(genediff, order(-r.pb, Names)), ]

     index<-1:p
     set<-matrix(0, p, 1)
     for(f in 1:S)
        {
           set[(f*5-4):(f*5), ]<-f
        }
     geneset<-data.frame(index, set)
     order_r<-order(r.pb, decreasing=T)

     phit<-matrix(rep(NA, S*p), nrow=S)
     pmiss<-matrix(rep(NA, S*p), nrow=S)
     for (s in 1:S)
        {
          for (i in 1:p)
               {
                  # calculate phit
                  hitindex<-intersect(geneset$index[geneset$set==s], order_r[1:i])
                   phit[s,i]<-sum(abs(r.pb[hitindex]))/
                                 sum(abs(r.pb[geneset$index[geneset$set==s]]))
```

```
                    # calculate pmiss
                  missindex<-intersect(geneset$index[geneset$set!=s], order_r[1:i])
                   pmiss[s,i]<-0
                   for (j in 1:length(missindex))
                       {
                          pmiss[s,i]<-pmiss[s,i]+
              (1/(p-length(geneset$set[geneset$set==geneset$set[missindex[j]]])))
                       }
                 }
          }

    es<-matrix(0, S, 1)
    for(k in 1:S)
        {
          d<-phit-pmiss
          es[k]<-if(max(d[k, ])== max(abs(d[k, ]))){max(d[k, ])} else {min(d[k, ])}
        }
    ES<-rbind(ES, es)
  }

ESS<-matrix(rep(NA, S*M), nrow=S)
for (v in 1:S)
  {
    for (w in 1:M)
        {
           ESS[v,w]<-ES[S*w-S+v, ]
        }
  }
```

```
#### compute P-Value ####

        pvalue.GSEA<-matrix(0, S, 1)
        for(t in 1:S)
            {
                pvalue.GSEA[t]<-if(ES.obs[t, ]>0){mean(ESS[t, ]>=ES.obs[t, ])}
                                    else {mean(ESS[t, ]<=ES.obs[t, ])}
                pvalue.GSEA<-as.matrix(pvalue.GSEA)
            }
        P_value.GSEA[aa, ]<-pvalue.GSEA[ ,1]

############################################################
############## Type I Error ###################

       type1.e.G<-0
       type1.e.L<-0

             tmp1<-sum(pvalue.GSEA[4:60]<0.05)
                    { if(tmp1>0) type1.e.G<-type1.e.G+1 }

             tmp2<-sum(pvalue.Logistic[4:60]<0.05)
                    { if(tmp2>0) type1.e.L<-type1.e.L+1 }

        Type1Error.GSEA[aa, ]<-type1.e.G
        Type1Error.Logistic[aa, ]<-type1.e.L

############## Power of Test ###################

             pow.G<-length(pvalue.GSEA[1:3][pvalue.GSEA[1:3]<0.05])
             pow.L<-length(pvalue.Logistic[1:3][pvalue.Logistic[1:3]<0.05])
```

```
        Power.GSEA[aa, ]<-pow.G/3
        Power.Logistic[aa, ]<-pow.L/3

       aa <- aa + 1
       if(aa > iterate) {break}
     }
}

P_value.GSEA<-data.frame(P_value.GSEA=P_value.GSEA)
write.table(P_value.GSEA, file = "P_value.GSEAc8_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

P_value.Logistic<-data.frame(P_value.Logistic=P_value.Logistic)
write.table(P_value.Logistic, file = "P_value.Logisticc8_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

TypeI.Error<-data.frame(Type1Error.GSEA= Type1Error.GSEA, Type1Error.Logistic=
Type1Error.Logistic)
write.table(TypeI.Error, file = "Type1Errorc8_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")

FWER.GSEA <- nnzero(as.numeric(Type1Error.GSEA), na.counted = FALSE)/iterate
FWER.Logistic <- nnzero(as.numeric(Type1Error.Logistic), na.counted = FALSE)/iterate
FWER<-data.frame(FWER.GSEA=FWER.GSEA, FWER.Logistic=FWER.Logistic)
write.table(FWER, file = "FWERc8_final.csv", sep = ",", col.names = TRUE,qmethod =
"double")

power<-data.frame(Power.GSEA= Power.GSEA, Power.Logistic= Power.Logistic)
ave.power.GSEA<-mean(power[ ,1])
sd.power.GSEA<-sd(power[ ,1])
ave.power.Logistic<-mean(power[ ,2])
```

```
sd.power.Logistic<-sd(power[ ,2])
stats.power<-data.frame(ave.power.GSEA, sd.power.GSEA, ave.power.Logistic,
sd.power.Logistic)

write.table(power, file = "powerc8_final.csv", sep = ",", col.names = TRUE,qmethod =
"double")
write.table(stats.power, file = "statspowerc8_final.csv", sep = ",", col.names =
TRUE,qmethod = "double")
```

ประวัติผู้เขียนวิทยานิพนธ์

นายสุธิภาส สิงห์เรือง เกิดวันจันทร์ที่ 15 ตุลาคม พ.ศ. 2533 สำเร็จการศึกษาปริญญาเศรษฐศาสตรบัณฑิต (ศ.บ.) เกียรตินิยมอันดับสอง (วิชาเอก : เศรษฐศาสตร์ปริมาณวิเคราะห์, วิชาโท : คณิตศาสตร์) คณะเศรษฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557