

การวิเคราะห์การถดถอยโลจิสติกส์แบบเอ็กแซ็ก สำหรับงานวิจัยที่มีตัวอย่างขนาดเล็ก: ตัวอย่างการวิเคราะห์ด้วยภาษาอาร์

ณัฐนารี เอมยงค์

อาจารย์ ภาควิชาชีวสถิติ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยมหิดล

บทคัดย่อ

การวิเคราะห์การถดถอยโลจิสติกส์เป็นวิธีที่ถูกนำไปประยุกต์ใช้ในการวิจัยด้านการแพทย์และสาธารณสุข รวมถึงด้านสุขศึกษาและพฤติกรรมศาสตร์ การประมาณค่าพารามิเตอร์ของตัวแบบอยู่ภายใต้ทฤษฎีแอสซิมโทติกซึ่งมีขนาดตัวอย่างที่มากพอ อย่างไรก็ตามตัวแบบที่มีขนาดตัวอย่างน้อย มีการกระจายสูง มีความเบ้ และตัวแปรตามมีการแยกกันอย่างชัดเจนก่อให้เกิดตัวประมาณที่ได้มีความเอนเอียง หรือไม่สามารถคำนวณค่าประมาณได้ ดังนั้นการวิเคราะห์การถดถอยโลจิสติกส์แบบเอ็กแซ็กจึงถูกนำมาเข้ามาแก้ปัญหาวิธีการประมาณค่าวิธีเอ็กแซ็กมีหลายรูปแบบโดยโปรแกรมภาษาอาร์ใช้ลูกโซ่ของมาร์คอฟในการประมาณค่าการถดถอยโลจิสติกส์แบบเอ็กแซ็ก บทความวิชาการนี้ได้แสดงวิธีการวิเคราะห์ข้อมูลด้วยภาษาอาร์และเปรียบเทียบกับวิธีมาตรฐาน

คำสำคัญ: การวิเคราะห์การถดถอยโลจิสติก/ ตัวอย่างขนาดเล็ก/ ภาษาอาร์

Exact Logistic Regression for Small Sample Size Research: An Example by R Language

Natnaree Aimyong

Lecturer of Department of Biostatistics, Faculty of Public Health, Mahidol University

Abstract

Logistic regression has been applied to medical and public health included health education and health behaviors researches. The estimation of model parameter needs the asymptotic theory which required large sample size. However, the small sample size, sparse, skew and completely separation cause the bias or the model cannot converge. Exact logistic regression has been implemented in those situations. There are many procedures for exact method, R language use Markov Chain Monte Carlo for estimating exact logistic regression. This academic article presented data analysis and comparison of results between exact and conventional logistic regression.

Keywords: Logistic regression/ Small sample size/ R language

บทนำ

การวิจัยด้านสุขศึกษาและพฤติกรรมศาสตร์ได้มีการประยุกต์ใช้การวิเคราะห์ถดถอยโลจิสติกส์เพื่อหาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามที่มีการวัดแบ่งออกเป็นสองกลุ่ม (dichotomous) โดยตัวแปรต้นสามารถเป็นได้ทั้งตัวแปรแบบกลุ่มและตัวแปรแบบต่อเนื่อง หรือเมื่อต้องการทำนายโอกาสเกิดตัวแปรตามจากตัวแปรต้น เช่น การศึกษาปัจจัยที่เกี่ยวข้องกับความแตกต่างด้านสุขภาพ พบว่าผู้สูงอายุ การศึกษาน้อย รายได้น้อย รู้สึกว่าไม่สุขภาพไม่ดี และการขาดการเข้าถึงอินเทอร์เน็ตมีโอกาสเป็นผู้มีความแตกต่างด้านสุขภาพต่ำเป็นต้น¹ การวิเคราะห์ข้อมูลด้วยการวิเคราะห์การถดถอยโลจิสติกส์นั้นการประมาณค่าพารามิเตอร์ของสมการใช้วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood estimator) ซึ่งอยู่ภายใต้ทฤษฎี Asymptotic ซึ่งต้องการขนาดตัวอย่างที่มีขนาดใหญ่ในการประมาณค่า² อย่างไรก็ตามในบางครั้งการวิจัยด้านสุขศึกษาและพฤติกรรมศาสตร์อาจมีการวิเคราะห์ในการวิจัยที่มีขนาดตัวอย่างน้อย

ในการศึกษาความแตกต่างด้านสุขภาพของนักศึกษาที่มีเชื้อสายละตินอเมริกา (Hispanic)³ การศึกษามีขนาดตัวอย่างทั้งสิ้น 331 คน แต่มีผู้ที่มีความแตกต่างด้านสุขภาพอยู่ในระดับต่ำเพียง 32 คน ผู้วิจัยใช้สมการถดถอยโลจิสติกส์เชิงเดียวพบว่า ตัวแปรภาษาที่ใช้ในครอบครัวมีความสัมพันธ์กับระดับความแตกต่างด้านสุขภาพ จากสถิติเชิงพรรณนาของงานวิจัยพบว่ามึนักศึกษาเชื้อสายละตินอเมริกาที่มีความแตกต่างด้านสุขภาพในระดับต่ำมีเพียง 3 คนที่ใช้ทั้งภาษาอังกฤษ และภาษาสเปนที่บ้าน

โดยนักศึกษาที่ไม่ใช้ภาษาอังกฤษและสเปน มี 1 คนที่มีความแตกต่างด้านสุขภาพอยู่ในระดับดี และไม่มีผู้ที่มีความแตกต่างอยู่ในระดับต่ำ ดังนั้นข้อมูลนี้จึงประกอบด้วยเซลล์ที่มีค่าสังเกตเป็นศูนย์ การวิจัยนี้แม้ว่าจะมีตัวอย่างขนาดใหญ่ แต่การที่มีเซลล์ที่มีค่าเป็นศูนย์ หรือมีค่าคาดหวังน้อยกว่า 5 มากกว่า 20% จะถือว่าข้อมูลมีปัญหาเรื่องขนาดตัวอย่างในการวิเคราะห์ข้อมูลด้วยการถดถอยโลจิสติกส์และอาจเกิดความเอนเอียงในการประมาณค่า

นอกจากนั้น ในการศึกษาหาปัจจัยที่เกี่ยวข้องกับพฤติกรรมการใช้ผลิตภัณฑ์ที่ทำให้ผิวขาวในนักศึกษาระดับปริญญาตรีในมหาวิทยาลัยแห่งหนึ่ง⁴ โดยมีตัวแปรตามคือการใช้ผลิตภัณฑ์ มีตัวแปรต้น ได้แก่ สาขาที่ศึกษา แบ่งเป็นสาขาวิทยาศาสตร์การแพทย์ และสาขาวิชาอื่นๆ ศึกษาในตัวแปรเพศโดยแบ่งออกเป็นชาย หญิง และเพศทางเลือก พบว่ามีจำนวนผู้ที่เป็นเพศทางเลือกทั้งสิ้น 10 คน โดยเป็นนักศึกษาในสาขาวิทยาศาสตร์การแพทย์เพียง 1 คน ซึ่งเมื่อควบคุมตัวแปรเพศและสาขาที่เรียนจะทำให้มีขนาดตัวอย่างไม่เพียงพอในการวิเคราะห์เช่นกัน

การวิเคราะห์ถดถอยโลจิสติกส์ในงานวิจัยที่มีขนาดเล็กนั้น มิได้ส่งผลต่อการประมาณค่าพารามิเตอร์เท่านั้น แต่ยังส่งผลถึงการวิเคราะห์เพื่อหาว่าตัวแบบที่สร้างขึ้นมีความเหมาะสมกับข้อมูลหรือที่เรียกว่าการวิเคราะห์สารูปดี (Goodness of fit) อีกด้วย วิธีการทางสถิติที่ใช้ในการทดสอบสารูปดีสำหรับการวิเคราะห์ถดถอยโลจิสติกส์ที่นิยม ได้แก่ การทดสอบเพียร์สันไคส์แควร์ (Pearson chi-square goodness of fit) การทดสอบดีไวเอนส์ (Deviance or likelihood ratio chi-square) ซึ่งเป็นการวัด

ความแตกต่าง และสัดส่วนระหว่างค่าสังเกตและค่าคาดหวังตามลำดับ การทดสอบสารูปดีด้วยวิธีการของเพียร์สัน ไคสแควร์ และดีเวียนส์อยู่ภายใต้สมมติฐานที่ว่าความแตกต่างของค่าสังเกตและค่าคาดหวังต้องเป็นแบบสุ่ม (Random) หรือไม่มีรูปแบบที่แน่นอนของความแตกต่างระหว่างค่าสังเกตและค่าคาดหวัง นอกจากนั้นวิธีการทดสอบสารูปดีทั้งสองวิธียังอยู่ภายใต้ทฤษฎี Asymptotic ดังนั้นวิธีการทั้งสองจะมีประสิทธิภาพลดลงเมื่อขนาดตัวอย่างมีขนาดเล็ก โดยมีแนวทางในการพิจารณาว่าตัวอย่างมีขนาดเล็กหรือไม่ดังนี้⁵

1. จำนวนตัวอย่างค่าสังเกตในแต่ละกลุ่มมีขนาดอย่างน้อย 10 คน
2. ร้อยละ 80 ของค่าคาดหวังมีค่าอย่างน้อย 5
3. เซลล์ที่มีค่าคาดหวังน้อยกว่า 5 ต้องมีค่าไม่น้อยกว่า 2 และไม่มีเซลล์ที่มีค่าเป็น 0

ดังนั้นเมื่อตัวแปรต้นเป็นตัวแปรที่มีการวัดแบบต่อเนื่อง หรือมีตัวแปรปฏิกริยาร่วม (Interaction) เป็นผลให้เซลล์ของตัวแปรเหล่านี้มักมีขนาดตัวอย่างขนาดเล็ก Hosmer และ Lemeshow⁶ ได้คิดวิธีแก้ปัญหานี้ด้วยการแบ่งกลุ่มของตัวอย่างออกเป็น 10 กลุ่ม (Deciles) ตามความน่าจะเป็นของตัวแปรตาม และทำนายจากตัวแบบ (Model-predicted probability) โดยใช้วิธีการเพียร์สัน ไคสแควร์เพื่อทดสอบสารูปดีของตัวแบบกับข้อมูลที่เป็นค่าสังเกต การแบ่งค่าความน่าจะเป็นออกเป็น 10 กลุ่มช่วยแก้ปัญหานี้ขนาดตัวอย่างที่มีขนาดน้อยในแต่ละเซลล์ในการทดสอบสารูปดีได้ อย่างไรก็ตาม การ

ทดสอบสารูปดี หรือการทดสอบความเหมาะสมของตัวแบบกับข้อมูลจะแสดงว่ามีความเหมาะสมสามารถสรุปได้ว่าไม่มีข้อมูลเชิงประจักษ์ว่าตัวแบบไม่เหมาะสมกับข้อมูล แต่ไม่สามารถบอกได้ถึงระดับความไม่เหมาะสมของตัวแบบนั้นๆ ต่อข้อมูล (lack of fit) ซึ่งสามารถทดสอบด้วยวิธี Pearson residual หรือวิธี Deviance residual⁷

จากปัญหาของการวิเคราะห์สมการถดถอยโลจิสติกส์ที่มีขนาดตัวอย่างขนาดเล็กข้างต้นนี้ มิได้มีปัญหต่อการทดสอบสารูปสนิทธิเท่านั้น แต่ยังส่งผลต่อการประมาณค่าพารามิเตอร์ของตัวแบบอีกด้วย การประมาณค่าพารามิเตอร์ของการวิเคราะห์การถดถอยโลจิสติกส์นั้นจะใช้ภาวะน่าจะเป็นสูงสุดภายใต้ทฤษฎีแอกซิโอมิติก ซึ่งเมื่อขนาดตัวอย่างมีขนาดน้อยอาจก่อให้เกิดความผิดพลาดในการประมาณค่าจากการที่การคำนวณไม่ลู่เข้าสู่ค่าประมาณ หรือไม่สามารถประมาณค่าได้ ในกรณีที่ตัวประมาณไม่ลู่เข้า (Converge) สู่อค่าประมาณ ถ้าผู้วิเคราะห์ข้อมูลไม่มีพื้นทางด้านสถิติอาจไม่ได้สังเกตว่าค่าประมาณที่โปรแกรมทางสถิติรายงานเรื่องของการไม่ลู่เข้าสู่ค่าประมาณ ซึ่งเป็นค่าประมาณที่ไม่สามารถนำไปใช้ได้ ในกรณีที่พบว่าสถิติเชิงพรรณนาแสดงถึงการที่ตัวอย่างมีขนาดเล็ก จึงแนะนำให้ใช้วิธี Exact logistic regression

วิธี Exact logistic regression สามารถใช้ในการแก้ปัญหา ขนาดตัวอย่างน้อย ตัวอย่างที่มีความเบ้ (Skew) ข้อมูลที่มีการกระจายสูง (Sparse) และข้อมูลที่ตัวแปรตามมีการแบ่งแยกกันอย่างเด็ดขาด (Complete separation)⁸ ข้อมูลที่มีลักษณะดังกล่าวมักเกิดปัญหาการไม่ลู่เข้าสู่ค่าประมาณ (Non-convergence) เมื่อวิเคราะห์

ข้อมูลด้วยการวิเคราะห์ถดถอยโลจิสติกส์แบบมาตรฐาน การใช้วิธี Exact logistic regression มีข้อเสียที่การวิเคราะห์ข้อมูลวิธีนี้ต้องการทรัพยากรด้านคอมพิวเตอร์มากและใช้เวลาในการวิเคราะห์นาน อย่างไรก็ตาม ด้วยความก้าวหน้าทางด้านคอมพิวเตอร์ที่มีประสิทธิภาพมากขึ้นจึงช่วยลดปัญหาเรื่องนี้ไปได้ ในปัจจุบันการวิเคราะห์ข้อมูลด้วยวิธี Exact logistic regression สามารถทำได้โดยใช้โปรแกรมทางสถิติ เช่น โปรแกรมอาร์ SAS และ Stata เป็นต้น

ตัวอย่างการวิเคราะห์ข้อมูล

ตัวอย่างการวิเคราะห์นี้มาจากผลการศึกษาของปพนสรค์ กลั่นเรืองแสง⁴ โดยตัวแปรตาม ได้แก่ การใช้ผลิตภัณฑ์เพื่อเพิ่มความขาวใสของผิว ตัวแปรต้น ได้แก่ ตัวแปรเพศ ซึ่งประกอบไปด้วย เพศหญิง เพศชายและเพศทางเลือก ตัวแปรสาขาที่ศึกษา ประกอบด้วยสาขาทางวิทยาศาสตร์การแพทย์ และไม่ใช่วิทยาศาสตร์การแพทย์ ผลการศึกษาพบว่ามีเซลล์ที่มีค่าศูนย์ และมีเซลล์ที่มีค่าคาตหมายน้อยกว่า 5 อยู่ 33% ดังนั้น การใช้การถดถอยโลจิสติกส์อาจเกิดความไม่เหมาะสม

ตารางที่ 1 รายละเอียดข้อมูลจำแนกตามตัวแปรต้นและตัวแปรตาม

กลุ่มที่	เพศ	สาขาที่เรียน	การใช้ผลิตภัณฑ์	จำนวน
1	เพศทางเลือก	วิทยาศาสตร์การแพทย์	ใช้ผลิตภัณฑ์	0
2	เพศทางเลือก	วิทยาศาสตร์การแพทย์	ไม่ใช้ผลิตภัณฑ์	1
3	เพศทางเลือก	สาขาอื่นๆ	ใช้ผลิตภัณฑ์	4
4	เพศทางเลือก	สาขาอื่นๆ	ไม่ใช้ผลิตภัณฑ์	5
5	ชาย	วิทยาศาสตร์การแพทย์	ใช้ผลิตภัณฑ์	12
6	ชาย	วิทยาศาสตร์การแพทย์	ไม่ใช้ผลิตภัณฑ์	25
7	ชาย	สาขาอื่นๆ	ใช้ผลิตภัณฑ์	2
8	ชาย	สาขาอื่นๆ	ไม่ใช้ผลิตภัณฑ์	37
9	หญิง	วิทยาศาสตร์การแพทย์	ใช้ผลิตภัณฑ์	31
10	หญิง	วิทยาศาสตร์การแพทย์	ไม่ใช้ผลิตภัณฑ์	91
11	หญิง	สาขาอื่นๆ	ใช้ผลิตภัณฑ์	15
12	หญิง	สาขาอื่นๆ	ไม่ใช้ผลิตภัณฑ์	96

วิธีการวิเคราะห์ข้อมูลด้วยวิธี Exact logistic regression

แม้มีโปรแกรมวิเคราะห์สถิติหลายตัวที่สามารถวิเคราะห์ Exact logistic regression ได้ ในที่นี้จะกล่าวถึงการวิเคราะห์ด้วยภาษาอาร์

(R language)⁹ เพราะเป็นฟรีแวร์ที่ทุกคนสามารถเข้าถึงได้โดยไม่มีค่าใช้จ่าย การใช้โปรแกรมอาร์เริ่มจากการดาวน์โหลดโปรแกรมจาก <http://CRAN.R-project.org/> โดยสามารถพิมพ์คำสั่งหลังเครื่องหมาย ">" การนำข้อมูลเข้า

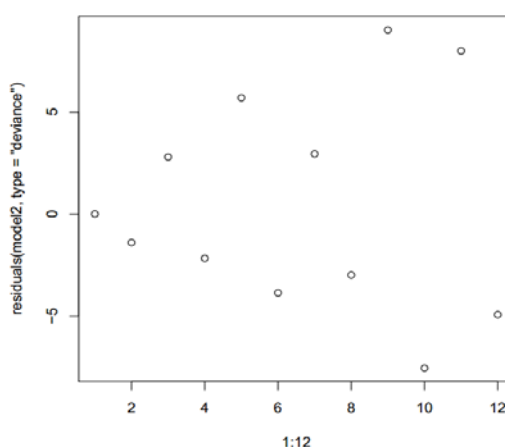
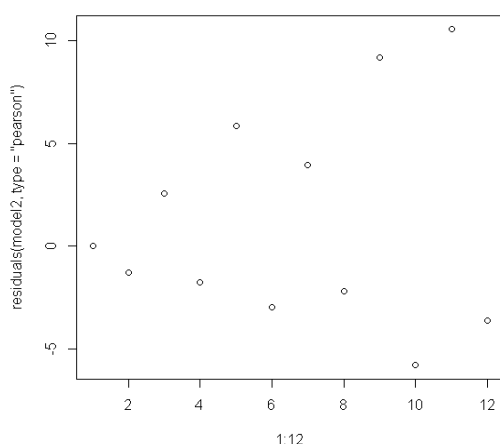
เพื่อวิเคราะห์ข้อมูลสามารถทำได้โดยการ save ข้อมูลให้เป็น .csv (comma separated value) โดยสามารถ save as จากโปรแกรมเอ็กเซลล์ จากนั้นนำข้อมูลเข้าด้วยคำสั่งดังรายละเอียดในภาคผนวก

การวิเคราะห์การถดถอยโลจิสติกส์ด้วยภาษาอาร์ สามารถทำได้โดยใช้คำสั่ง “glm” ในตัวอย่างนี้สมการโลจิสติกส์มีตัวแปรต้นสองตัวได้แก่ เพศ (sex) และ สาขาที่เรียน (major) ซึ่งต้องสั่งให้โปรแกรมทราบว่าตัวแปรนี้เป็นตัวแปรเชิงกลุ่ม ด้วยคำสั่ง factor() และระบุว่าเป็นการวิเคราะห์การถดถอยโลจิสติกส์ด้วยคำสั่ง family=binomial("logit") โดยระบุว่าจะให้ใช้ data frame ชื่อ d ด้วยคำสั่ง data=d

จากผลการวิเคราะห์ซึ่งแสดงในตารางที่ 2 พบว่า เพศและสาขาที่เรียนมีความสัมพันธ์กับการใช้ผลิตภัณฑ์เพื่อเพิ่มความขาวใส โดยเพศหญิงและเพศชายมีการใช้ผลิตภัณฑ์ไม่แตกต่างกัน ($OR = \exp(-0.05) = 1.05$, $95\%CI = (-0.76, 0.60)$) เพศ

ทางเลือกมีความเสี่ยงการใช้ผลิตภัณฑ์สูงกว่าเพศหญิง 4.4 เท่า ($\exp 4.49$) โดยมีค่า $95\%CI (0.05, 2.9)$ นักศึกษาสาขาวิทยาศาสตร์การแพทย์ใช้ผลิตภัณฑ์ 2.7 เท่าเมื่อเทียบกับนักศึกษาสาขาอื่นๆ $95\%CI (0.42, 1.63)$

ข้อมูลชุดนี้มีปัญหาเรื่องขนาดตัวอย่างเล็ก จึงทดสอบสารูปสนิทธิด้วย Hosmer-Lemeshow test ผลการทดสอบพบว่าไม่มีข้อมูลเชิงประจักษ์ว่าตัวแบบไม่เหมาะสมกับข้อมูล ($p=0.9983$) จากนั้นจึงทำการทดสอบระดับความไม่เหมาะสมกับตัวแบบ (lack of fit) จากผลการทดสอบความไม่เหมาะสมกับตัวแบบตั้งแผนภูมิที่ 1 พบว่า กลุ่มเพศหญิงที่ใช้ผลิตภัณฑ์ (กลุ่มที่ 9 และ 11) ซึ่งมีค่า pearson residual และ deviance residual ที่มีค่าสูง โดยปกติอาจแก้ไขด้วยการเพิ่ม interaction ลงไปในตัวแบบ อย่างไรก็ตาม ในข้อมูลนี้อาจเกิดจากจำนวนตัวอย่างขนาดเล็ก ดังนั้นจึงควรวิเคราะห์ข้อมูลด้วย exact logistic regression



แผนภูมิที่ 1 ผลการทดสอบระดับความไม่เหมาะสมของตัวแบบ (lack of fit) ด้วยวิธี Pearson residual (ซ้าย) และ Deviance (ขวา)

ในการวิเคราะห์ Exact logistic regression ภาษาอาร์ใช้คำสั่ง “elrm”¹⁰ ตัวแปรตามรูปแบบการเขียนคำสั่งนี้จะป็นในรูปแบบ จำนวนครั้งที่สำเร็จต่อจำนวนทั้งหมด เช่น ในตัวอย่างนี้มีตัวอย่างที่เป็นเพศทางเลือกและเรียนในสาขา

วิทยาศาสตร์การแพทย์ทั้งหมด 9 คน มี 4 คนที่ใช้ผลิตภัณฑ์เพื่อเพิ่มความขาวใส หรือมีค่าตัวแปรตามเป็น 4/9 สำหรับกลุ่มย่อยนี้ เป็นต้น จึงต้องมีการจัดการข้อมูลก่อนทำการวิเคราะห์ดังรายละเอียดในภาคผนวก

ตารางที่ 2 การเปรียบเทียบผลการวิเคราะห์ด้วยการถดถอยโลจิสติกส์ และการถดถอยเอ็กแซ็กโลจิสติกส์

วิธีวิเคราะห์	ตัวแปร	ค่าพารามิเตอร์	ช่วงความเชื่อมั่นที่ 95%	
			ขอบจำกัดล่าง	ขอบจำกัดบน
การถดถอยโลจิสติกส์แบบมาตรฐาน	หญิง	1		
	ชาย	-0.052	-0.756	0.604
	เพศทางเลือก	1.494	0.049	2.853
	สาขาวิทยาศาสตร์การแพทย์	1		
	สาขาอื่นๆ	1.003	0.415	1.626
การถดถอยเอ็กแซ็กโลจิสติกส์	หญิง	1		
	ชาย	-0.086	-0.833	0.602
	เพศทางเลือก	0.985	-0.616	2.487
	สาขาวิทยาศาสตร์การแพทย์	1		
	สาขาอื่นนอกจากวิทยาศาสตร์การแพทย์	0.085	0.267	1.530

จากตารางที่ 2 ผลการวิเคราะห์ พบว่าค่าพารามิเตอร์จากการประมาณค่าจากการถดถอยโลจิสติกส์แบบมาตรฐานมีค่าสูงกว่าวิธี Exact logistic regression ยกเว้นตัวแปรเพศชาย นอกจากนั้นค่าพารามิเตอร์ของตัวแปรเพศทางเลือกมีความแตกต่างกันมากระหว่างการวิเคราะห์สองวิธีนี้ ดังนั้นวิธีการวิเคราะห์ถดถอยโลจิสติกส์แบบมาตรฐานมีแนวโน้มของการ over estimate เมื่อเปรียบเทียบกับถดถอยเอ็กแซ็กโลจิสติกส์ในกรณีนี้พบว่าขนาดตัวอย่างมีขนาดเล็ก ดังนั้นจึงแนะนำให้ใช้ผลการวิเคราะห์จากการถดถอยเอ็กแซ็กโลจิสติกส์

สรุป

การวิเคราะห์การถดถอยโลจิสติกส์นั้นมีการประมาณค่าพารามิเตอร์ภายใต้ทฤษฎี Asymptotic ซึ่งต้องการขนาดตัวอย่างที่มากพอ ดังนั้นเมื่อตัวอย่างมีขนาดเล็ก การความเบ้ มีการกระจายสูง หรือตัวแปรตามมีการแบ่งแยกกันอย่างชัดเจน จึงควรวิเคราะห์ข้อมูลด้วย วิธี Exact logistic regression เพื่อลดความผิดพลาดจากการประมาณค่า อย่างไรก็ตามข้อเสียของวิธีนี้คือต้องการทรัพยากรคอมพิวเตอร์สูง การถดถอยเอ็กแซ็กโลจิสติกส์สามารถวิเคราะห์ด้วยโปรแกรมทางสถิติ

หลายโปรแกรม โดยโปรแกรมภาษาอาร์เป็นโปรแกรมที่เป็นฟรีแวร์ที่สามารถเข้าถึงได้โดยไม่มีค่าใช้จ่าย

ภาคผนวก

เวอร์ชันของโปรแกรมอาร์ที่ใช้ได้แก่ R 3.0.1 โดยวิธีวิเคราะห์ด้วยวิธีของ Bruin¹¹ ตัวอักษรหลังเครื่องหมาย '#' เป็นคำอธิบายไม่สามารถรับคำสั่งได้

ให้โปรแกรมอ่านข้อมูลชื่อ exact.csv มาสู่โปรแกรมอาร์ใน data frame ชื่อ d

```
>d=read.csv("setwd("C:/Documents and Settings/Natnaree/Desktop/exact.csv")
```

สั่งให้โปรแกรมแสดงตัวแปรทั้งหมดใน data frame ชื่อ d

```
> ls(d)
```

```
[1] "major" "SEX" "Use" # ผลลัพธ์คำสั่งแสดงตัวแปรทั้งหมดใน data frame ชื่อ d ในกรณีนี้มีตัวแปร 3 ตัวได้แก่ major SEX และ Use
```

การวิเคราะห์การถดถอยโลจิสติกส์ด้วยวิธีมาตรฐาน $Y_{Use} = X_{Sex} + X_{major}$ เมื่อ X_{Sex} และ X_{major} เป็นตัวแปรเชิงกลุ่ม โดยให้เก็บผลการวิเคราะห์ในข้อมูลชื่อ model

```
>Model = glm(Use~factor(SEX)+factor(major), data=d, family=binomial("logit"))
```

```
> summary(model) # แสดงผลการวิเคราะห์
```

วิเคราะห์สารรูปสถิติด้วยวิธีของ Hosmer and Lemanshow

```
> library(ResourceSelection)
```

```
> hoslem.test( model$y,fitted(model))
```

ระดับความไม่เหมาะสมของตัวแบบ (lack of fit) ซึ่งต้องมีการคำนวณข้อมูลตามจำนวนกลุ่มดังนี้

```
>tab1=as.data.frame(table(d$major,d$SEX,d$Use))
```

```
>colnames(tab1)[1] <- 'major'
```

```
>colnames(tab1)[2] <- 'sex'
```

```
>colnames(tab1)[3] <- 'use'
```

```
>model2=glm(use~factor(sex)+factor(major), data=tab1, weight= Freq, family="binomial")
```

สร้างกราฟระดับความไม่เหมาะสมของตัวแบบด้วย pearson residual และ deviance residual

```
>plot(1:12, residuals(model2,type="pearson"))
```

```
>plot(1:12, residuals(model2,type="deviance"))
```

การวิเคราะห์สมการถดถอยเอ็กแซ็กโลจิสติกส์

```
>install.packages("elrm")
```

```
>library(elrm)
```

```
>tab2 = xtabs(~Use + interaction(sex, major), data = d)
```

```
>xdat = data.frame(sex = rep(0:2, 2), major = rep(0:1, each = 3), Use = x[2, ], ntrials = colSums(tab2))
```

วิเคราะห์และแสดงผลของตัวแปรเพศ และแสดงผล

```
>s = elrm(formula = Use/ntrials ~ factor(sex), interest = ~factor(sex), iter = 500000,
```

```
dataset = xdat, burnIn = 50000)
```

```
summary(s)
```

วิเคราะห์ แสดงผลของตัวแปรสาขาที่เรียนและแสดงผล

```
m <- elrm(formula = Use/ntrials ~ major, interest = ~major, iter = 500000,
```

```
dataset = xdat, burnIn = 50000)
```

```
summary(m)
```

เอกสารอ้างอิง

1. Protheroe J, Whittle R, Bartlam B, Estacio EV, Clark L, Kurth J. Health literacy, associated lifestyle and demographic factors in adult population of an English city: a cross-sectional survey. *Health Expectations* 2017 Feb 1; 20(1): 112-9.
2. Kleinbaum DG, Klein M, Pryor ER. Logistic regression: a self-learning text. 3rd ed. 2010.
3. Mas S, Jacobson H, Dong Y. Health literacy level of Hispanic college students. *Southern medical journal* 2014 Feb 1; 107(2): 61-5.
4. Paponson Klanruangsang. Behavior of selecting Skin Whitening products: A case study among undergraduate students in Bangkok. [Master of Science (Biostatistics)]. Bangkok: Mahidol University; 2017.
5. Stokes ME, Davis CS, Koch GG. Categorical data analysis using SAS. SAS institute; 2012 July 31.
6. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*. 1997 May 15; 16(9): 965-80.
7. Kleinbaum D, Kupper L, Nizam A, Muller K. Applied regression analysis and other multivariable methods. Cengage Learning; 2007 Apr 23.
8. Wilson JR, Lorenz KA. Modeling binary correlated responses using SAS, SPSS and R. Springer; 2015 Oct 12.
9. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria [Internet]. 2007. [cited 15 Dec 2017]; Available from: (<https://www.R-project.org/>).
10. Zamar D, McNeney B, Graham J. Elrm: Software implementing exact-like inference for logistic regression models. *Journal of Statistical Software* 2007 Oct 8; 21(3): 1-8.
11. Bruin J. Newtest: command to compute new test [Internet]. 2011. [cited 15 Dec 2017]. Available from: (<https://stats.idre.ucla.edu/stata/ado/analysis/>).