NU Science Journal 2007; 3(2): 171 - 187

การวิเคราะห์ข้อมูลขนาดสมองและปัจจัยที่เกี่ยวข้องภายใต้ตัวแบบโลจิท เมื่อตัวแปรอธิบายเป็นแบบเชิงกลุ่มทั้งหมดและไม่ใช่เชิงกลุ่มทั้งหมด

อวิกา โรจน์วิรัตน์ และ วีรานันท์ พงศาภักดี*

The Analysis of Brain Size under the Logit Model for Categorical Explanatory Data and Mixed Data

Awika Rojwiratana and Veeranun Pongsapukdee*

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

*Corresponding author. E-mail: veeranun@su.ac.th

บทคัดย่อ

การวิเคราะห์ข้อมูลขนาดสมองและปัจจัยที่เกี่ยวข้องโดยใช้ตัวแบบโลจิทศึกษาข้อมูลจากเว็บไซต์ http://lib.stat.cmu.edu/ ซึ่งเป็นข้อมูลขนาดของสมองกับเชาว์ปัญญาของ Willerman และคณะ (1991) ที่สำรวจมาจากนักศึกษาจิตวิทยาเบื้องต้นที่ถนัดมือขวาของมหาวิทยาลัยเซาท์เวสเทริน ประเทศอังกฤษ จำนวน 38 คน ที่ไม่มีประวัติการติดเหล้า เสียสติ มีปัญหาทางสมอง หรือเป็นโรคหัวใจ วัตถุประสงค์ของการวิจัย เพื่อศึกษาอิทธิพลของตัวแปรอธิบายแบบต่าง ๆ ที่มีผลต่อตัวแปรตอบสนองขนาดสมอง (Y) ที่เป็นแบบสองกลุ่ม คือ กลุ่มที่มีขนาดสมองน้อยกว่า 900,000 pixels กับกลุ่มที่มีขนาดสมองมากกว่าหรือเท่ากับ 900,000 pixel โดยตัวแปรอธิบายได้แก่ เพศ (Gender) คะแนน IQ รวม (FSIQ) คะแนน IQ ในเรื่องการพูด (VIQ) คะแนน IQ ในเรื่องการแสดงออก (PIQ) น้ำหนัก (Weight) และส่วนสูง (Height) การวิเคราะห์ข้อมูลแบ่งออกเป็น 2 กรณี คือ กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่มและกรณีที่ข้อมูลของตัวแปรอธิบายไม่เป็นแบบเชิงกลุ่ม ทั้งสองกรณีนี้ประมวลผลภายใต้ตัวแบบโลจิทด้วยวิธีการคัดเลือกตัวแปรแบบขั้นตอน (Stepwise) โดยใช้โปรแกรมสำเร็จรูป SAS (Version 9.1)

ผลการวิจัยพบว่า ปัจจัยที่มีอิทธิพลต่อขนาดของสมองอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 คือ เพศ และคะแนน IQ ในเรื่องการแสดงออก นอกจากนี้ยังพบว่าตัวแบบโลจิทกรณีที่ตัวแปรอธิบายเป็นแบบเชิงกลุ่ม $(\text{logit}(\hat{P}(x)) = -4.003 + 4.581\text{Gender} - 1.774\text{PIQ} - \text{Group})$ มีความเหมาะสมมากกว่าตัวแบบ $(\text{logit}(\hat{P}(x)) = -0.349 + 4.560\text{Gender} - 0.063\text{PIQ})$ ซึ่งเป็นตัวแบบโลจิทกรณี

ที่ตัวแปรอธิบายไม่เป็นเชิงกลุ่ม เนื่องจากตัวแบบโลจิทกรณีแรกที่ตัวแปรอธิบายเป็นแบบเชิงกลุ่มให้ค่าของอัตราส่วนภาวะน่าจะเป็น (G^2) AIC และ BIC น้อยกว่าตัวแบบโลจิทกรณีที่ตัวแปรอธิบายไม่เป็นเชิงกลุ่ม และให้ค่าวัดความสัมพันธ์และความสอดคล้องแบบ Somers'D, Gamma และ c ได้มากกว่าตัวแบบโลจิทกรณีที่ข้อมูลของตัวแปรอธิบายไม่เป็นเชิงกลุ่ม ตลอดจนให้ค่าส่วนเหลือที่น้อยกว่าเป็นส่วนใหญ่และมีการกระจายที่สวยงามกว่ากรณีหลังอีกด้วย

คำสำคัญ : ตัวแบบโลจิท ตัวแปรตอบสนองแบบมีลำดับ ตัวแปรอธิบายเชิงกลุ่ม

Abstract

The data and the involved factors of this analysis came from website http://lib.stat.cmu.edu/ on "Brain Size Data and Intelligence" which belongs to Willerman et al. (1991). They collected a sample of 38 right-handed Ango introductory psychology students at Southwestern University. The students had indicated no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease. The research was to study the effect of explanatory variables on the dichotomous response of the student's brain size, where one group had less than 900,000 pixels and the other had more than or equal to 900,000 pixels. The explanatory variables were gender, full scale IQ scores (FSIQ), verbal IQ scores (VIQ), performance IQ scores (PIQ), weight and height. The analysis of data was based on logit models that consisted of two cases: the grouped explanatory variables and the ungrouped explanatory variables. The selection of models under the stepwise method was performed using SAS version 9.1.

The results revealed that the significant effect factors at $\alpha = 0.05$ to the brain size of the students were Gender and Performance IQ scores (PIQ). The logit model in the case of categorical explanatory variable ($\text{logit}(\hat{P}(x)) = -4.003+4.581\text{Gender}-1.774\text{PIQ\_Group}$) was preferable to the logit model in the case of ungrouped explanatory variable ($\text{logit}(\hat{P}(x)) = 0.034+4.560\text{Gender}-0.063\text{PIQ\_Group}$). The likelihood ratio (G^2), AIC and BIC statistics of the first case were less than those of the second case. Moreover, Somers'D, Gamma, and c statistics of the first case were also outperformed, including the magnitude of residuals. The residual plots of both cases were found independently scattered.

Keywords: Logit Models, Dichotomous, Categorical Explanatory

บทนำ

การวิเคราะห์ตัวแบบโลจิท (Logit model) มีตัวแปรที่สนใจหลักคือตัวแปรตอบสนอง (Response variable) เป็นตัวแปรเชิงสุ่ม (Random variable) แบบเชิงกลุ่ม (Categorical)ที่มีมาตรวัดแบบนามบัญญัติ (Nominal) หรือแบบมีลำดับ (Ordinal) และมีตัวแปรอธิบาย (Explanatory variables) ซึ่งโดยทั่วไปเป็นตัวแปรแบบเชิงกลุ่มหรืออาจมีแบบต่อเนื่องด้วย ตัวแปรตอบสนอง เช่น การสอบได้หรือ สอบตก การเป็นโรคชนิดหนึ่งแบบรุนแรง ปานกลาง หรือไม่เป็นโรค ผลการดำเนินการหลักทรัพย์ ดีมาก ดี ปานกลาง และปิดกิจการ ฯลฯ ซึ่งเป็นข้อมูลที่พบได้ทั่วไปและสามารถประยุกต์ใช้แพร่หลายในสาขาวิชาต่าง ๆ เช่น วิทยาศาสตร์สุขภาพ วิศวกรรมศาสตร์ สังคมศาสตร์ ฯลฯ ตัวแปรตอบสนองเชิงกลุ่มนี้อาจแบ่งออกเป็น 2 กรณีคือ กรณีที่ตัวแปรตอบสนองเป็นแบบสองกลุ่ม (Binary or dichotomous variable) โดยกำหนดให้ตัวแปรตอบสนองมีค่าเท่ากับ 1 เมื่อเกิดเหตุการณ์ที่สนใจในหน่วยที่ i และมีค่าเท่ากับ 0 หรือ 2 เมื่อไม่เกิดเหตุการณ์ที่สนใจในหน่วยที่ i และกรณีที่ตัวแปรตอบสนองเป็นแบบหลายกลุ่มหรือมากกว่าสองกลุ่ม (Polychotomous variable) เช่น ระดับการศึกษาสูงกว่าปริญญาตรี ปริญญาตรี และต่ำกว่าปริญญาตรี รายได้ สูง กลาง ต่ำ ฯลฯ

งานวิจัยนี้สนใจศึกษากรณีที่ตัวแปรตอบสนองเป็นแบบสองกลุ่ม ที่เกี่ยวกับขนาดของสมองที่น้อยกว่า 900,000 pixels $\left(Y=1\right)$ และขนาดของสมองที่มากกว่าหรือเท่ากับ 900,000 pixels $\left(Y=2\right)$ ของนักศึกษามหาวิทยาลัยเซาท์เวสเทริน ประเทศอังกฤษ (Willerman *et al*., 1991) โดยจำแนกกลุ่มของการเกิดเหตุการณ์ที่สนใจ และที่ไม่สนใจให้มีค่าของตัวแปรเป็น 1 และ 2 ตามลำดับ การวิเคราะห์ใช้รูปแบบของความน่าจะเป็น (Probability) หรือ $P\left(x\right)$ แทนความน่าจะเป็นของ $Y=1$ เมื่อกำหนด X โดยที่ X แทนปัจจัยที่ส่งผลต่อ Y ซึ่งโดยทั่วไปนิยมใช้การแปลงโลจิทให้อยู่ในรูปโลจิทของ $P\left(x\right)$ คือ Logit ของ $Q\left(x\right)=\log\left[\frac{P\left(x\right)}{1-P\left(x\right)}\right]$ ในรูปแบบดังนี้

$$\log\left[\frac{P\left(x\right)}{1-P\left(x\right)}\right]=\beta_0+\beta_1 x$$

ตัวแบบโลจิท คือ Logarithm ของ Odds ในรูปความสัมพันธ์เชิงเส้นกับ X โดยที่ $P\left(x\right)=P\left(Y=1\mid x\right)$ งานวิจัยนี้นอกจากสนใจการนำตัวแบบโลจิทมาอธิบายขนาดสมองที่เป็นแบบสองกลุ่มแล้ว ยังสนใจศึกษาปัจจัยที่ส่งผลต่อ Y หรือตัวแปรอธิบายที่มีนัยสำคัญทางสถิติที่เน้นการวิเคราะห์เป็นสองกรณีคือ กรณีที่ข้อมูลของตัวแปรอธิบายไม่เป็นแบบเชิงกลุ่ม (Ungrouped explanatory variables) หมายถึงข้อมูลที่เป็นค่าของตัวแปรอธิบายต่าง ๆ ของหน่วยทดลองโดยตรงที่มีทั้งแบบเชิงกลุ่มและแบบต่อเนื่อง และกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม (Categorical or grouped explanatory variables) หมายถึงข้อมูลที่มีการแปลงค่าของตัวแปรอธิบายต่าง ๆ ให้เป็น

แบบเชิงกลุ่มเพียงอย่างเดียวโดยสามารถจัดอยู่ในรูปแบบของตารางการณ์จร(Contingency table) ซึ่งเป็นทางเลือกในการวิเคราะห์ข้อมูล เนื่องจากการแจกแจงของตัวสถิติที่ใช้วัดภาวะสารูปดี (Goodness-of-fit) ของตัวแบบเช่นตัวสถิติเพียรสัน (Pearson statistics) และตัวสถิติอัตราส่วนภาวะน่าจะเป็น (Likelihood ratio statistic) ภายใต้สมมติฐานว่างนั้น มีการแจกแจงแบบไคสแควร์ ที่มีองศาอิสระเท่ากับ J-(p+1) หรือ χ^2_{J-p-1} เมื่อ J แทนจำนวนรูปแบบหรือจำนวนกลุ่มของตัวแปรอธิบาย (Covariate pattern) ข้อเสียของตัวสถิติทั้งสองนี้ คือ เมื่อไม่มีการจัดกลุ่มหรือเมื่อ J ≈ n (ขนาดตัวอย่าง) การแจกแจงภายใต้ข้อมูลแต่ละหน่วยที่ไม่จัดกลุ่มหรือมีขนาด n คือ χ^2_{n-p-1} ซึ่งจำนวนพารามิเตอร์จะเพิ่มขึ้นตามขนาดของตัวอย่าง กรณีเช่นนี้ทำให้ตัวสถิติทั้งสองข้างต้นภายใต้การแจกแจง χ^2_{n-p-1} จะให้ค่าของ p-value ที่ไม่ถูกต้อง Hosmer และ Lemeshow (1989) McCullagh และ Nelder (1983) ได้แสดงว่าตัวสถิติทั้งสองดังกล่าวมีค่าคาดหวังน้อยกว่าองศาอิสระ J-(p+1) ด้วย หรือที่เรียกว่ามีปัญหาของ n-asymptotics ซึ่งทางหนึ่งที่จะเหลีกเลี่ยงปัญหานี้คือการจัดกลุ่มข้อมูลเพื่อให้แต่ละกลุ่มสอดคล้องกับ m-asymptotics เมื่อ m=J<n และวิธีการแก้ปัญหาหนึ่งคือการกำหนดจำนวนกลุ่มที่แน่นอน m กลุ่ม เมื่อ m<J โดยอาจรวมกลุ่มที่มีความถี่คาดหวังน้อย (Lemeshow and Hosmer, 1982)

ดังนั้นงานวิจัยนี้จึงสนใจการวิเคราะห์เป็นสองกรณี คือ กรณีที่ข้อมูลของตัวแปรอธิบายไม่เป็นแบบเชิงกลุ่มและแบบเชิงกลุ่ม โดยใช้ตัวแบบโลจิทภายใต้วิธีภาวะน่าจะเป็นสูงสุด เพื่อศึกษาปัจจัยที่ส่งผลหรือมีอิทธิพลต่อตัวแปรตอบสนอง ข้อมูลจริงที่ใช้เผยแพร่ใน http://lib.stat.cmu.edu เกี่ยวกับขนาดของสมองกับเชาวน์ปัญญาของนักศึกษาจิตวิทยาเบื้องต้นที่ถนัดมือขวาของมหาวิทยาลัยเซาท์เวสเทริน ประเทศอังกฤษ ซึ่งจำแนกขนาดสมองออกเป็น 2 กลุ่มที่มีขนาดของสมองน้อยกว่า 900,000 pixels และมากกว่าหรือเท่ากับ 900,000 pixels แทนตัวแปรตอบสนอง และศึกษาตัวแปรอธิบายที่ส่งผลหรือมีอิทธิพลต่อตัวแปรตอบสนองจากตัวแบบโลจิท โดยการเลือกตัวแปรแบบขั้นตอน (Stepwise selection) การตรวจสอบภาวะสารูปดี ของตัวแบบใช้ตัวสถิติอัตราส่วนภาวะน่าจะเป็น (G^2), AIC (Akaike's information criterion), BIC (Bayesian information criterion) และร้อยละของการพยากรณ์ถูกต้อง ตลอดจนตัวสถิติการวัดความสัมพันธ์ของความน่าจะเป็นของการพยากรณ์และค่าสังเกตด้วยตัวสถิติ Somers'D, Gamma และ c (concordant)

ตัวอย่างและขอบเขตของการวิจัย

ข้อมูลตัวอย่างมาจากนักศึกษาจิตวิทยาเบื้องต้นที่ถนัดมือขวาของมหาวิทยาลัยเซาท์เวสเทริน ประเทศอังกฤษ ที่ไม่มีประวัติการติดเหล้า เสียสติ มีปัญหาทางสมอง หรือเป็นโรคหัวใจ โดยใช้การแบ่งการทดสอบออกเป็น 4 ส่วนย่อย ๆ คือ คำศัพท์ (Vocabulary) ความเหมือน (Similarity) บล็อกดีไซน์ (Block design) และการต่อภาพ (Picture completion) ซึ่งเป็นการทดสอบที่เรียกว่า

The Wechsler (1981) Adult Intelligence Scale-Revised ของ Willerman และคณะ (1991) และใช้ Magnetic Resonance Imaging (MRI) ในการบ่งชี้ขนาดของสมองของคน รวมไปถึงข้อมูลเพศและขนาดของร่างกาย (ความสูงและน้ำหนัก) จำนวน 38 คน จากเว็บไซต์ http:// lib.stat.cmu.edu/โดยมีตัวแปรต่าง ๆ ดังนี้

1. ตัวแปรตอบสนองแบบ 2 กลุ่ม จำแนกด้วยขนาดของสมองที่น้อยกว่าหรือเท่ากับและมากกว่า 900,000 pixels ที่วัดด้วยเครื่อง MRI เป็นกลุ่ม 1 และ 2 ตามลำดับ

2. ตัวแปรอธิบายที่ใช้ในงานวิจัยนี้มีทั้งหมด 6 ตัวแปร ได้แก่ เพศ (Gender) คะแนน IQ โดยรวม (Full scale IQ scores) คะแนน IQในเรื่องของการพูด (Verbal IQ scores) คะแนน IQ ในเรื่องของการแสดงออก (Performance IQ scores) น้ำหนัก (Weight) และความสูง (Height) โดยทำการศึกษาเป็น 2 กรณี คือกรณีที่ข้อมูลของตัวแปรอธิบายที่ไม่เป็นแบบเชิงกลุ่ม (Ungrouped explanatory variable) และกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม (Categorical or grouped explanatory variable) โดยแปลงค่าของตัวแปรอธิบายแบบต่อเนื่องให้เป็นเชิงกลุ่มแบบมีลำดับโดยใช้เกณฑ์ที่ผู้วิจัยกำหนดขึ้นและจำแนกข้อมูลตามระดับหรือกลุ่มให้อยู่ในรูปตารางการณ์จร

3. การเลือกตัวแปรอธิบายเข้าในตัวแบบใช้วิธีการเลือกตัวแปรแบบขั้นตอน (Stepwise selection)

4. การวิเคราะห์ข้อมูลใช้ระดับความมีนัยสำคัญที่ระดับ 0.05 และไม่คำนึงถึงอันดับของตัวแปร

5. การประมวลผลใช้โปรแกรม SAS (Statistical analysis systems) Version 9.1

การดำเนินการวิจัย

งานวิจัยนี้แบ่งข้อมูลเป็น 2 กรณี คือ กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม คือใช้ข้อมูลจริงของแต่ละหน่วยทดลองที่เก็บรวบรวมโดยตรง ซึ่งประกอบด้วยตัวแปรอธิบายทั้งที่เป็นตัวแปรเชิงกลุ่มและตัวแปรต่อเนื่อง และกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่มโดยแปลงข้อมูลของตัวแปรอธิบายแบบต่อเนื่องให้มีลักษณะเป็นตัวแปรเชิงกลุ่มแบบมีลำดับดังแสดงไว้ในตาราง 1-2 และประมวลผลข้อมูลโดยใช้โปรแกรม SAS ซึ่งให้ผลลัพธ์เป็น 2 กรณีตามลำดับ

1. กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม (Ungrouped explanatory variables)

ตาราง 1 ตัวแปรอธิบายและตัวแปรตอบสนองเมื่อข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

ตัวแปรอธิบาย	คำอธิบาย
Gender	เพศ : Gender = 1 แทนนักศึกษาชาย และ Gender = 2 แทนนักศึกษาหญิง
FSIQ	คะแนน IQ โดยรวม ที่วัดจากการทดสอบย่อยของ Wechsler (1981)
VIQ	คะแนน IQ ในเรื่องของการพูด ที่วัดจากการทดสอบย่อยของ Wechsler (1981)
PIQ	คะแนน IQ ในเรื่องของการแสดงออก ที่วัดจากการทดสอบย่อยของ Wechsler (1981)
Weight	น้ำหนักของร่างกาย (ปอนด์)
Height	ความสูงของร่างกาย (นิ้ว)

2. กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม (Categorical explanatory variables)

ตาราง 2 ตัวแปรอธิบายและตัวแปรตอบสนองเมื่อข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

ตัวแปร	คำอธิบาย
Gender	เพศ : Gender = 1 แทนนักศึกษาชาย และ Gender = 2 แทนนักศึกษาหญิง
FSIQ_Group	คะแนน IQ โดยรวม ที่วัดจากการทดสอบย่อยของ Wechsler (1981)
	FSIQ = 1 นักศึกษาที่สอบได้คะแนน IQ โดยรวม $\leq$ 103 คะแนน
	FSIQ = 2 นักศึกษาที่สอบได้คะแนน IQ โดยรวม $>$ 103 คะแนน
VIQ_Group	คะแนน IQ ในเรื่องของการพูด ที่วัดจากการทดสอบย่อยของWechsler (1981)
	VIQ = 1 นักศึกษาที่สอบได้คะแนน IQ ในเรื่องของการพูด $<$ 103 คะแนน
	VIQ = 2 นักศึกษาที่สอบได้คะแนน IQ ในเรื่องของการพูดอยู่ในช่วง 103-130 คะแนน
	VIQ = 3 นักศึกษาที่สอบได้คะแนน IQ ในเรื่องของการพูด $>$ 130 คะแนน

ตาราง 2 (ต่อ)

ตัวแปร	คำอธิบาย
PIQ_Group	คะแนน IQ ในเรื่องของการแสดงออก ที่วัดจากการทดสอบย่อยของ Wechsler (1981)
	PIQ = 1 นักศึกษาที่สอบได้คะแนน IQในเรื่องของการแสดงออก < 103 คะแนน
	PIQ = 2 นักศึกษาที่สอบได้คะแนนIQในเรื่องของการแสดงออกในช่วง 103–130 คะแนน
	PIQ = 3 นักศึกษาที่สอบได้คะแนน IQในเรื่องของการแสดงออก > 130 คะแนน
Weight_Group	น้ำหนักของร่างกาย (ปอนด์)
	Weight = 1 นักศึกษาที่มีน้ำหนัก < 135 ปอนด์
	Weight = 2 นักศึกษาที่มีน้ำหนักอยู่ในช่วง 135 – 165 ปอนด์
	Weight = 3 นักศึกษาที่มีน้ำหนัก > 165 ปอนด์
Height_Group	ความสูงของร่างกาย (นิ้ว)
	Height = 1 นักศึกษาที่มีความสูง < 65 นิ้ว
	Height = 2 นักศึกษาที่มีความสูงอยู่ในช่วง 65 – 75 นิ้ว
	Height = 3 นักศึกษาที่มีความสูง > 75 นิ้ว
ตัวแปรตอบสนอง	
MRI_Count	ขนาดของสมองที่วัดจากเครื่อง Magnetic Resonance Imaging : MRI (pixels)
	MRI_Count = 1 นักศึกษาที่มีขนาดของสมองน้อยกว่า 900,000 pixels
	MRI_Count = 2 นักศึกษาที่มีขนาดของสมองมากกว่าหรือเท่ากับ 900,000 pixels

การวิเคราะห์ข้อมูล

กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม มี 3 ขั้นตอนดังนี้

1. การคัดเลือกตัวแปรแบบขั้นตอนโดยใช้ตัวสถิติอัตราส่วนภาวะน่าจะเป็น (Likelihood ratio: LR) ซึ่งมีขั้นตอนคล้ายการวิเคราะห์แบบขั้นตอนในการวิเคราะห์การถดถอยแบบพหุ โดยจำนวนตัวแปรที่อยู่ในตัวแบบพิจารณาจากค่า Score statistics ของตัวแปรอธิบายแต่ละตัว ซึ่งค่า Score statistics มาจากการคำนวณเมตริกซ์ร่วมกันคือ เมตริกซ์ของการทำอนุพันธ์บางส่วนของฟังก์ชันภาวะน่าจะเป็นลำดับที่ 1 และลำดับที่ 2 ซึ่งแต่ละเมตริกซ์มาจากการประมาณด้วยวิธีภาวะน่าจะเป็น ตัวแปร

อิสระใดที่มีค่า Score Statistics มากที่สุดจะถูกนำเข้าตัวแบบเป็นอันดับแรก และทำการเพิ่มจำนวนตัวแปรอธิบายทีละตัวโดยพิจารณาจากค่า Score statistics ที่มีค่ามากรองลงมาตามลำดับ

2. ตัวแปรอธิบายที่คัดเลือกได้ในข้อ 1 นำมาประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และพิจารณาความมีนัยสำคัญของแต่ละตัวแปรจากสถิติของ Wald

3. ทำการวัดความสัมพันธ์ของความน่าจะเป็นของการพยากรณ์และค่าสังเกตด้วยตัวสถิติ Somers' D, Gamma และ c และการตรวจสอบภาวะสารูปดี ของตัวแบบว่าตัวแบบใดมีความเหมาะสมมากน้อยเพียงใด โดยใช้ตัวสถิติอัตราส่วนภาวะน่าจะเป็น (G^2) ตัวสถิติ AIC และตัวสถิติ BIC

กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

การวิเคราะห์กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม ทำในทำนองเดียวกันกับกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม ต่างกันเฉพาะลักษณะของข้อมูลในกรณีตัวแปรอธิบายเป็นแบบเชิงกลุ่มได้จากการนำตัวแปรอธิบายที่เป็นแบบต่อเนื่องมาแปลงให้มีลักษณะเป็นเชิงกลุ่มก่อน แล้วจึงนำมาวิเคราะห์หาตัวแบบ และวัดความสัมพันธ์ด้วยตัวสถิติ Somers'D, Gamma และ c และตรวจสอบภาวะสารูปดี ของตัวแบบด้วยตัวสถิติอัตราส่วนภาวะน่าจะเป็น (G^2) ตัวสถิติ AIC และตัวสถิติ BIC ตามลำดับ

ผลการวิจัย แบ่งเป็น 2 กรณี ดังนี้
กรณีที่ 1 ผลการวิเคราะห์กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม
กรณีที่ 2 ผลการวิเคราะห์กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

กรณีที่ 1 ผลการวิเคราะห์กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

จากตาราง 3 พบว่าค่าประมาณพารามิเตอร์ $\hat{\beta}$, Standard error ($S.E.\hat{\beta}$) และตัวสถิติ Wald ของตัวแบบโลจิทที่ได้จากการเลือกตัวแปรแบบขั้นตอน มีปัจจัยหรือตัวแปร 2 ตัวที่มีนัยสำคัญที่ 0.05 ที่ถูกเลือกเข้าในสมการ คือ Gender และ PIQ_Group และตัวแบบของ Logit ตัวอย่าง (Estimated logit) คือ

$$\text{logit}(\hat{P}(x)) = -0.349 + 4.560\text{Gender} - 0.063\text{PIQ}$$

ถ้ากำหนดให้ Gender = 2 และ PIQ = 124 จะได้ค่าของ $\text{logit}(\hat{P}(x)) = 0.959$ ค่าของโลจิทตัวอย่างที่ได้สามารถนำไปคำนวณหาค่าของ Odds ของ Y = 1 คือ exp(0.959) หรือเท่ากับ 2.609 และค่าของ $\hat{P}(x)$ ที่เท่ากับ $e^{0.959}/(1+e^{0.959})$ หรือเท่ากับ 0.723 ซึ่งค่าของ $\hat{P}(x)$ ที่ได้คือความน่าจะเป็น

ที่นักศึกษาจะเป็นกลุ่มที่ 1 $\left[P(Y=1|x)\right]$ หมายความว่าถ้านักศึกษาเป็นเพศหญิงและได้คะแนน IQ ในเรื่องการแสดงออกเท่ากับ 124 คะแนน ความน่าจะเป็นที่ขนาดสมองของนักศึกษาจะน้อยกว่า 900,000 pixel จะเท่ากับ 0.723 (ดูตาราง 3 ประกอบ)

ตาราง 3 ค่าประมาณพารามิเตอร์ของตัวแบบโลจิท กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

ตัวแปร	$\hat{\beta}$	$S.E.\hat{\beta}$	Wald	df	p-value
Intercept	-0.349	2.529	0.019	1	0.890
Gender	4.560	1.451	9.884*	1	0.002
PIQ	-0.063	0.032	3.943*	1	0.047

จากตาราง 4 พบว่าประสิทธิภาพในการพยากรณ์ของตัวแบบกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่มในภาพรวมสามารถพยากรณ์ค่าได้ถูกต้องถึง 81.6% เมื่อแยกพิจารณาประสิทธิภาพในการพยากรณ์เป็นส่วน ๆ พบว่า ถ้าขนาดสมองของนักเรียนน้อยกว่า 900,000 pixels (MRI = 1) จะพยากรณ์ได้ถูกต้อง 83.3% และถ้าขนาดของสมองมากกว่าหรือเท่ากับ 900,000 pixels (MRI = 2) จะพยากรณ์ได้ถูกต้อง 80.0% (ดูตาราง 4 ประกอบ)

ตาราง 4 ประสิทธิภาพในการพยากรณ์ กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

ค่าสังเกต		ค่าพยากรณ์ (Predicted Value)		
		MRI		% ของการพยากรณ์
		1	2	ถูกต้อง
MRI	1	15	3	83.3
	2	4	16	80.0
รวม		**19**	**9**	**81.6**

จากตาราง 5 ผลการทดสอบสัมประสิทธิ์ทุกตัวของตัวแบบเท่ากับ 0 หรือไม่ พบว่าตัวสถิติ Likelihood ratio, Score และ Wald ให้ผลลัพธ์เหมือนกัน ซึ่งผลที่ได้ชี้ให้เห็นว่าค่าสัมประสิทธิ์บางตัว

ในตัวแบบมีค่าไม่เท่ากับ 0 อย่างมีนัยสำคัญ กล่าวอีกนัยหนึ่งคือ ตัวแปร Gender และ PIQ ที่อยู่ในตาราง 3 นั้นมีความเหมาะสมที่จะใช้ในการพยากรณ์

ตาราง 5 ผลของการทดสอบสมมติฐานว่าง ($\beta = 0$) ของตัวแบบ กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

Test	Chi-Square	df	P-value
Likelihood Ratio	26.106*	2	< 0.0001
Score	20.791*	2	< 0.0001
Wald	10.019*	2	0.0067

จากตาราง 6 พบว่าร้อยละของจำนวนคู่ของค่าสังเกตที่มีความสอดคล้องกัน (Percent concordant) มากกว่าร้อยละของจำนวนคู่ของค่าสังเกตที่ไม่สอดคล้องกัน (Percent discordant) ดังนั้นตัวแปรตอบสนองและตัวแปรอธิบายมีความสัมพันธ์ในทิศทางเดียวกัน และเมื่อพิจารณาจากค่าของตัวสถิติ Somers'D, Gamma และ c พบว่าตัวสถิติดังกล่าวให้ค่าเป็นบวก โดยที่ตัวสถิติ Somers'D, Gamma และ c มีค่าของความสัมพันธ์เท่ากับ 0.839, 0.844 และ 0.919 ตามลำดับ ซึ่งชี้ให้เห็นว่ามีความสัมพันธ์ทางบวกค่อนข้างมากระหว่างความน่าจะเป็นของการพยากรณ์และค่าสังเกต

ตาราง 6 การวัดความสัมพันธ์ของความน่าจะเป็นของการพยากรณ์ค่าสังเกตกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

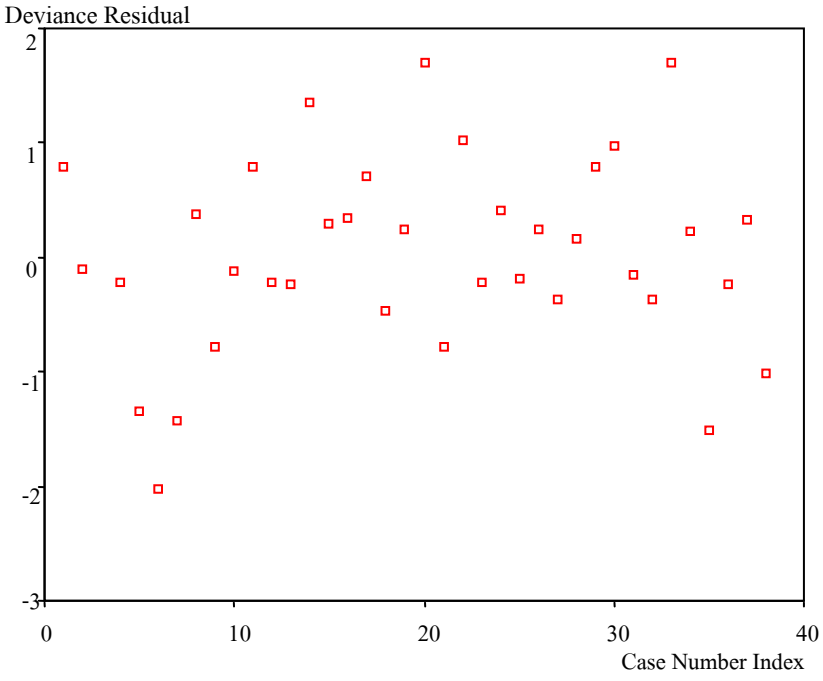
ตัวสถิติ	ค่าของความสัมพันธ์
Percent concordant	91.7
Percent discordant	7.8
Somers'D	0.839
Gamma	0.844
c	0.919

จากตาราง 7 พบว่าตัวแบบโลจิทมีความเหมาะสม โดยให้ค่าของตัวสถิติที่ใช้ในการวิเคราะห์ภาวะสารูปดี G^2 หรือ -2 log L มีค่า 26.468 ซึ่งมีค่าค่อนข้างต่ำ ส่วน AIC และ BIC เท่ากับ 32.468 และ 37.207 ตามลำดับ

ตาราง 7 ความเหมาะสมของตัวแบบ กรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

Criterion	Intercept and Covariates
-2 log L	26.468
AIC	32.468
BIC	31.207

จากรูป 1 เป็นการพล็อตการกระจายของส่วนเหลือ (Deviance Residuals) ตามลำดับกับค่าของตัวแปรตอบสนองกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม เพื่อตรวจสอบการกระจายรอบค่าศูนย์ พบว่าส่วนใหญ่แล้วค่า Deviance Residuals ของค่าสังเกตแต่ละค่านั้นมีการกระจายรอบ ๆ ศูนย์ และมีค่าอยู่ในช่วงที่ยอมรับได้ คือ มีค่าอยู่ระหว่าง $[-2,2]$ ดังนั้นตัวแบบที่นำมาใช้มีความสอดคล้องกับข้อมูล โดยมีส่วนเหลือต่ำ



รูป 1 การกระจายของ Deviance Residuals ตามลำดับค่าสังเกตของตัวแปรตอบสนองกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่ม

กรณีที่ 2 ผลการวิเคราะห์กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

จากตาราง 8 ค่าประมาณพารามิเตอร์ $\hat{\beta}$, Standard error ($S.E.\hat{\beta}$) และตัวสถิติ Wald ของตัวแบบโลจิทที่ได้จากการเลือกตัวแปรแบบขั้นตอน พบว่ามีปัจจัยหรือตัวแปรอธิบาย 2 ตัวที่ถูกเลือกเข้าในสมการเช่นเดียวกับของส่วนที่ 1 คือ Gender และ PIQ_Group ที่ ระดับนัยสำคัญ 0.05 และสามารถเขียนตัวแบบ Logit ตัวอย่าง (Estimated logit) ได้ดังนี้

$$\text{logit}(\hat{P}(x)) = -4.003 + 4.581\text{Gender} - 1.774\text{PIQ\_Group}$$

ถ้ากำหนดให้ Gender = 2 และ PIQ_Group = 2 ได้ค่าของ logit $(\hat{P}(x))$ = 1.611 ค่าของโลจิทตัวอย่างที่ได้นั้นสามารถนำไปคำนวณหาค่าของ Odds ของ Y = 1 คือ exp(1.611) หรือเท่ากับ 5.007 และค่าของ $\hat{P}(x)$ ที่เท่ากับ $e^{1.611}/(1+e^{1.611})$ หรือเท่ากับ 0.834 ซึ่งค่าของ $\hat{P}(x)$ ที่ได้คือความน่าจะเป็นที่นักศึกษาจะเป็นกลุ่มที่ 1 $\left[P(Y=1|x)\right]$ หมายความว่าถ้านักศึกษาเป็นเพศหญิงและได้คะแนน IQ ในเรื่องการแสดงออกอยู่ในช่วง 103-130 คะแนน ความน่าจะเป็นที่ขนาดสมองของนักศึกษาจะน้อยกว่า 900,000 pixels จะเท่ากับ 0.834

ตาราง 8 ค่าประมาณพารามิเตอร์ของตัวแบบโลจิท กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

ตัวแปร	$\hat{\beta}$	S.E.$\hat{\beta}$	Wald	Df	p-value
Intercept	-4.003	1.708	5.491	1	0.0191
Gender	4.581	1.393	10.810*	1	0.001
PIQ_Group	-1.774	0.815	4.738*	1	0.029

จากตาราง 9 พบว่าประสิทธิภาพในการพยากรณ์ของตัวแบบกรณีที่ข้อมูลเป็นแบบเชิงกลุ่มในภาพรวมสามารถพยากรณ์ค่าได้ถูกต้องถึง 81.1% เมื่อแยกพิจารณาประสิทธิภาพในการพยากรณ์เป็นส่วน ๆ พบว่า ถ้าขนาดสมองของนักเรียนน้อยกว่า 900,000 pixels (MRI = 1) พยากรณ์ได้ถูกต้อง 72.2% และถ้าขนาดของสมองมากกว่าหรือเท่ากับ 900,000 pixels (MRI = 2) พยากรณ์ได้ถูกต้อง 90%

ตาราง 9 ประสิทธิภาพในการพยากรณ์ กรณีที่ข้อมูลของตัวแปรตอบอธิบายเป็นแบบเชิงกลุ่ม

ค่าสังเกต		ค่าพยากรณ์ (Predicted Value)		
		MRI		ร้อยละของการ
		1	2	จำแนกถูก
MRI	1	13	5	72.2
	2	2	18	90.0
รวม		**15**	**23**	**81.1**

ตาราง 10 ผลการทดสอบค่าสัมประสิทธิ์ทุกตัวของตัวแบบ พบว่าตัวสถิติ Likelihood ratio score และ Wald ให้ผลลัพธ์เหมือนกัน คือค่าสัมประสิทธิ์บางตัวในตัวแบบมีค่าไม่เท่ากับ 0 อย่างมีนัยสำคัญ กล่าวอีกนัยหนึ่งคือ ตัวแปรที่ได้จากการคัดเลือกตัวแปรแบบขั้นตอนคือ Gender และ PIQ_Group ที่อยู่ในตาราง 8 นั้นมีความเหมาะสมที่จะใช้ในการพยากรณ์

ตาราง 10 ผลของการทดสอบสมมติฐานว่าง ($\beta = 0$) ของตัวแบบกรณีที่มีข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

Test	Chi-Square	df	P-value
Likelihood ratio	27.616*	3	< 0.0001
Score	21.735*	3	< 0.0001
Wald	10.391*	3	0.016

จากตาราง 11 ร้อยละของจำนวนคู่ของค่าสังเกตที่มีความสอดคล้องกัน (Percent concordant) พบว่ามากกว่าร้อยละของจำนวนคู่ของค่าสังเกตที่ไม่ความสอดคล้องกัน (Percent discordant) นั่นคือตัวแปรตอบสนองและตัวแปรอธิบายมีความสัมพันธ์ในทิศทางเดียวกัน ส่วนค่าของตัวสถิติ Somers'D, Gamma และ c พบว่าให้ค่าเป็นบวก โดย Somers'D Gamma และ c มีค่าเท่ากับ 0.856, 0.917 และ 0.928 ตามลำดับ ซึ่งชี้ให้เห็นถึงความสัมพันธ์ทางบวกในทิศทางเดียวกันค่อนข้างมากระหว่างตัวแปร

ตาราง 11 ความสัมพันธ์ของความน่าจะเป็นของการพยากรณ์และค่าสังเกต กรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

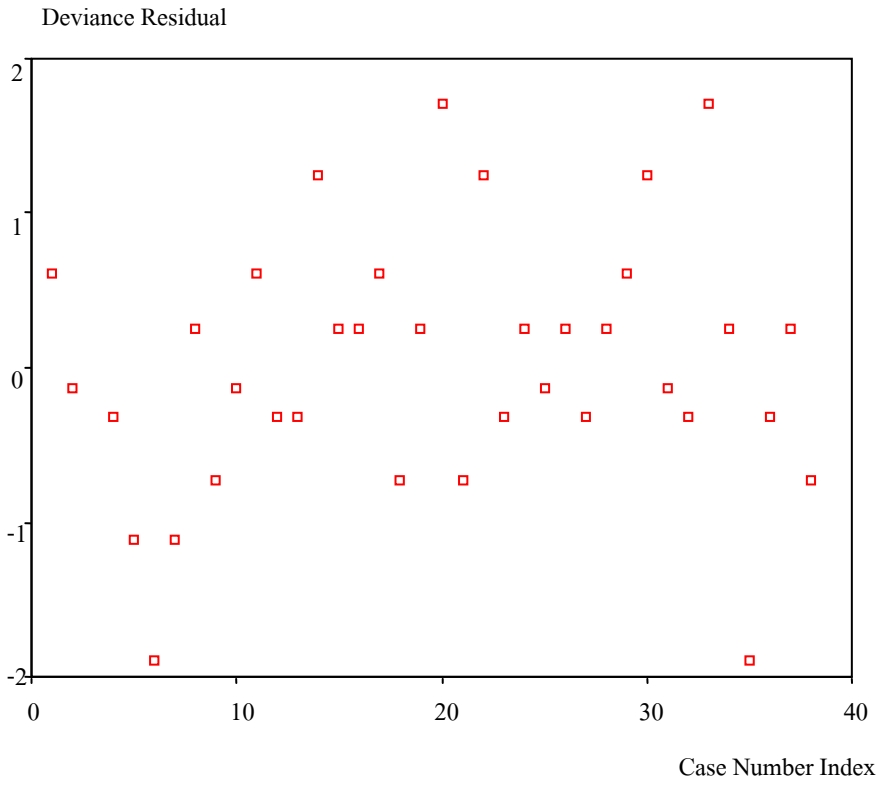
ตัวสถิติ	ค่าของความสัมพันธ์
Percent concordant	89.4
Percent discordant	3.9
Somers'D	0.856
Gamma	0.917
c	0.928

ตาราง 12 เมื่อตัวแปรตอบสนองเป็นแบบสองกลุ่มกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม พบว่าตัวแบบมีความเหมาะสม G^2 AIC และ BIC มีค่า 25.609, 31.609 และ 31.343 ตามลำดับ

ตาราง 12 ความเหมาะสมของตัวแบบ กรณีที่ข้อมูลข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

Criterion	Intercept and Covariates
-2 log L	25.609
AIC	31.609
BIC	30.348

รูป 2 เป็นการพล็อตการกระจายของ Deviance residual พบว่าส่วนใหญ่แล้วค่า Deviance residuals ของค่าสังเกตแต่ละค่านั้นมีการกระจายรอบศูนย์ และไม่มีรูปแบบ โดยมีค่าอยู่ในช่วงที่ยอมรับได้คือมีค่าอยู่ระหว่าง $[-2,2]$ ดังนั้นตัวแบบที่นำมาใช้มีความสอดคล้องกับข้อมูล



รูป 2 การกระจายของ Deviance residuals ตามลำดับค่าสังเกตของตัวแปรตอบสนองกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่ม

สรุปผลการวิจัยและข้อเสนอแนะ

จากผลการวิจัย พบว่าตัวแบบโลจิทที่ได้จากการวิเคราะห์ข้อมูลกรณีที่ข้อมูลของตัวแปรอธิบายผสมไม่เป็นแบบเชิงกลุ่มคือ

$$\text{logit}(\hat{P}(x)) = -0.349 + 4.560\text{Gender} - 0.063\text{PIQ} \qquad (1)$$

และตัวแบบโลจิทที่ได้จากการวิเคราะห์ข้อมูลกรณีที่ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่มคือ

$$\text{logit}(\hat{P}(x)) = -4.003 + 4.581\text{Gender} - 1.774\text{PIQ\_Group} \qquad (2)$$

ตัวแบบ (1) ให้ร้อยละของความถูกต้องในการพยากรณ์เท่ากับ 81.6 และให้ค่าอัตราส่วนภาวะน่าจะเป็น (G^2) = 26.468, AIC = 32.468 และ BIC = 31.207 ทั้งยังให้ค่าวัดความสัมพันธ์ Somers'D = 0.839, Gamma = 0.844 และ c = 0.919 อีกด้วย

ตัวแบบ (2) ให้ร้อยละของความถูกต้องในการพยากรณ์เท่ากับ 81.1 และให้ค่าอัตราส่วนภาวะน่าจะเป็น (G^2) = 25.609, AIC = 31.609 และ BIC = 30.348 ตลอดจนยังให้ค่าวัดความสัมพันธ์ Somers'D = 0.856, Gamma = 0.917 และ c = 0.928 อีกด้วย

การตีความหมายของตัวแบบทั้งสองตัวแบบคล้ายกัน พบว่า ตัวแปรที่ส่งผลต่อขนาดสมองมากที่สุดคือตัวแปรเพศ (Gender) และคะแนน IQ (PIQ_Group) เช่น ตัวแบบ (2) อัตราส่วน Odds ของขนาดสมองเป็นกลุ่มที่ 1 ของหญิงต่อชายเท่ากับ exp(4.581) และเมื่อคะแนน PIQ ลดลง 1 ระดับ มีผลให้อัตราส่วน Odds ของขนาดสมองเป็นกลุ่มที่ 1 เป็น 5.89 เท่าของคะแนน PIQระดับเดิม อีกทั้งตัวแบบ (2) ให้ร้อยละของความถูกต้องในการพยากรณ์ที่ใกล้เคียงกับตัวแบบ(1) แต่ยังให้ค่าค่าอัตราส่วนภาวะน่าจะเป็น (G^2) ค่า AIC และค่า BIC ที่ต่ำกว่า (อาจเหมาะสมกว่า) ตลอดจนให้ค่าวัดความสัมพันธ์ Somers'D, Gamma และ c ที่มากกว่า (เหมาะสมกว่า) ของตัวแบบ (1) ด้วย

ดังนั้นผู้วิจัยจึงอาจเลือกใช้ตัวแบบ (2) ซึ่งก็คือตัวแบบโลจิทของกรณีที่ใช้ข้อมูลของตัวแปรอธิบายเป็นแบบเชิงกลุ่มในกรณีที่ 2 ในการวิเคราะห์และพยากรณ์ข้อมูลต่อไป

อย่างไรก็ตาม งานวิจัยนี้เป็นผลสรุปจากการใช้ข้อมูลจริงเพียงชุดเดียว ผลของการวิจัยจึงเป็นข้อสังเกตที่น่าศึกษาเพิ่มเติมต่อไปได้อีก โดยเฉพาะในเรื่องการเลือกใช้วิธีวิเคราะห์วิธีใดภายใต้ข้อมูลแบบใดให้ชัดเจนขึ้น นอกจากนี้อาจจัดกลุ่มของขนาดสมอง (ตามผู้เชี่ยวชาญ) แบบมากกว่าสองกลุ่มแบบมีลำดับโดยอาจศึกษาตัวแบบใหม่ ๆ ที่เหมาะสมตามตัวแปรตอบสนองที่มากกว่า 2 กลุ่ม และเพิ่มการตรวจสอบตัวแบบเช่นการตรวจสอบความคงทน (Tolerance) ของตัวแบบและอื่นๆ ด้วย

เอกสารอ้างอิง

วีรานันท์ พงศาภักดี (2541). *การวิเคราะห์ข้อมูลเชิงกลุ่ม : ทฤษฎีและการประยุกต์* (กับ GLIM และ SPSS/FW) โรงพิมพ์มหาวิทยาลัยศิลปากร นครปฐม.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.) Second Edition, New York : John Willey and Sons, Inc.

Akaike, H. (1985). *Prediction and entropy*, In A Celebration of Statistic (eds A.C. Atkinson and S.E.Fienberg), Springer-Verlag, Berlin : 1-24.

Burnham, K.P. and Anderson, D.R.(1992). *Data-based selection of an appropriate biological model : the key to modern data analysis*, In Wildlife 1 : Populations (eds D.R. McCullough and R.H. Barrett), Elsevier Science Publishers, London : 16-30.

Burnham, K.P., Anderson, D.R. and White G.C. (1994). Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models, *Biom J*, *36,*299-315.

Cox, D.R. (1970).*Analysis of Binary Data*, New York : Chapman and Hall.

Cox, D.R. and Snell , E.J. (1989). *The Analysis of Binary Data* (2nd ed.) London : Chapman and Hall, 1989.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models, *J.Roy, Statest.Soc.Ser*, *Al35,* 370-384.

SAS Institute Inc. (1995). *Logistic Regression Examples Using the SAS System*, Cary , NC : SAS Institute Inc.

SAS Institute Inc. (2003). *SAS OnlineDoc®9.1 Cary*, NC : SAS Institute Inc.

Stokes , M.E. , Davis , C.S. and Koch , G.G. (2000). *Categorical Data Analysis Using the SAS System* (2nd ed.) Cary, NC : SAS Institute Inc.

Willerman, L., Schultz, R., Rutledge,J.N. and Bigler .E. (1991). *In Vivo* Brain Size and Intelligence, Intelligence, *15*, 223-228. http:// lib.stat.cmu.edu /