การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อตัวแปรตอบสนองมี 2 กลุ่ม

นางสาวกุลพัชร หมื่นมา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต

สาขาวิชาสถิติ ภาควิชาสถิติ

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2551

A Comparison of Parameter Estimation Methods for Binary Response of Logit, Probit, and Complementary log-log Models

Miss Kunlaphat Muenma

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 2008

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของ ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อตัวแปรตอบสนองมี 2 กลุ่ม
โดย	นางสาวกุลพัชร หมื่นมา
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

.... ประธานกรรมการ

(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

.... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร. บุญอ้อม โฉมที)

.... กรรมการ

(อาจารย์ ดร.อรุณี กำลัง)

.... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

กุลพัชร หมื่นมา : การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อตัวแปรตอบสนองมี 2 กลุ่ม. (A Comparison of Parameter Estimation Methods for Binary Response of Logit, Probit, and Complementary log-log Models) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา, 111 หน้า.

งานวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบ วิธีการประมาณค่าพารามิเตอร์ใน ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก วิธีการประมาณค่าพารามิเตอร์ที่ใช้ในงานวิจัยครั้งนี้ คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS), วิธีภาวะน่าจะเป็นสูงสุด (MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) โดยที่ตัวแปรตอบสนอง ทั้ง 3 วิธี เป็นตัวแปรเชิงคุณภาพมี 2 ค่า คือ 0 หรือ 1 และ ตัวแปรอธิบาย(X) 1 ตัวแปร การเปรียบเทียบกระทำภายใต้ข้อมูลงานทดลองของ Draper(1972), Ashford(1970), Cornfield (1962), Martin(1942), Muhammad(1990), (Strand,1930), Montgomery(1982), Clogg (1988) และ Haberman(1978) ตัวอย่างชุดที่ 1-3 เป็นข้อมูลทางด้านการแพทย์ ตัวอย่างชุดที่ 4-6 เป็นข้อมูลทางด้านวิทยาศาสตร์(ชีววิทยา) ตัวอย่างชุดที่ 7 เป็นข้อมูลทางด้านวิศวกรรมศาสตร์ และตัวอย่างชุดที่ 8-9 เป็นข้อมูลทางด้านสังคมศาสตร์ วิธีการวิเคราะห์ข้อมูล คือ การประมาณค่าของตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ซึ่งใช้ตัวสถิติ Deviance เป็นเกณฑ์ในการเปรียบเทียบ โดยใช้โปรแกรม R ผลการวิเคราะห์ข้อมูลพบว่า

1. ตัวแบบโลจิทภายใต้วิธี MLE ให้ค่า Deviance น้อยสุด เป็นตัวแบบที่เหมาะสมกับข้อมูลตัวอย่างชุดที่ 4
2. ตัวแบบโพรบิทภายใต้วิธี MLE ให้ค่า Deviance น้อยสุด เป็นตัวแบบที่เหมาะสมกับข้อมูลตัวอย่างชุดที่ 2 ชุดที่ 5 และ ชุดที่ 9
3. ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MLE ให้ค่า Deviance น้อยสุด เป็นตัวแบบที่เหมาะสมกับข้อมูลตัวอย่างชุดที่ 3 และ ชุดที่ 7
4. สำหรับข้อมูลชุดที่ 1 ชุดที่ 6 และ ชุดที่ 8 ไม่มีตัวแบบใดเหมาะสมกับข้อมูล เพราะความน่าจะเป็นในการยอมรับตัวแบบน้อยกว่าระดับนัยสำคัญ 0.05

ในส่วนการพิจารณาการเลือกตัวแบบให้มีความเหมาะสมกับลักษณะของข้อมูล 9 ชุด นั้นจะเห็นว่าขนาดของความแปรปรวนขึ้นอยู่กับขนาดของข้อมูล และ ความแตกต่างของค่าพารามิเตอร์ที่ประมาณได้จะขึ้นอยู่กับขนาดของ σ เท่านั้น ดังนั้น ตัวแปรอธิบายมีอิทธิพลต่อโอกาสความน่าจะเป็นของการเกิดสิ่งที่สนใจ $P(y=1)$ ซึ่งมีผลต่อการเลือกลิงค์ฟังก์ชัน

ภาควิชา สถิติ ลายมือชื่อนิสิต กุลพัชร หมื่นมา
สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก
ปีการศึกษา 2551

4982250026 : MAJOR STATISTICS

KEY WORD: LOGIT MODEL / PROBIT MODEL / COMPLEMENTARY LOG-LOG MODEL / MAXIMUM LIKELIHOOD ESTIMATION / WEIGHTED LEAST SQUARE / MINIMUM CHI-SQUARE

KUNLAPHAT MUENMA : A COMPARISON OF PARAMETER ESTIMATION METHODS FOR BINARY RESPONSE OF LOGIT, PROBIT, AND COMPLEMENTARY LOG-LOG MODELS THESIS PRINCIPAL ADVISOR : ASSOC.PROF.SUPOL DURONGWATANA,Ph.D., 111 pp.

The objective of this research is to compare three methods of parameter estimation in logit model, probit model, and complementary log-log model. These methods are Weighted Least Squares Estimation (WLS), Maximum Likelihood Estimation (MLE) and Minimum Chi-Square Estimation (MCS). Response variables of all three models are binary variables with one explanatory variable (X). The comparison are investigated under nine sets of sample data appeared in Draper(1972), Ashford(1970), Cornfield (1962), Martin(1942), Muhammad(1990), (Strand,1930), Montgomery(1982), Clogg (1988) and Haberman(1978). The sample 1-3 are medical science data, sample 4-6 are biological science data, sample 7 is engineering data and sample 8-9 are social science data. The three models are compared by employing the deviance as the performance measure . All data are analyzed using R statistical package.

1. Logit model fitted by MLE method yields the smallest deviance for the sample 4.
2. Probit model fitted by MLE method yields the smallest deviance for the sample 2, 5 and 9.
3. Complementary log-log model fitted by MLE method yields the smallest deviance for the sample 3 and 7.
4. There are the unsuitable model for the sample 1, 6 and 8 because the probability of accept model less than 0.05 significance.

In the past of determination of model selection for nine sample data, size of the variance is affected by size of sample and the difference parameter estimation is only affected by size of σ . So that the probability of success $P(y=1)$ is affected by explanatory variable and it results in selection of link function.

Department : Statistics Student's signature : Kunlaphat Muenma
Field of study : Statistics Principal Advisor's signature :
Academic year : 2008

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลุล่วงไปด้วยความช่วยเหลืออย่างดียิ่งของ รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำ ปรึกษา ตลอดจนช่วยเหลือแก้ไขข้อบกพร่องต่างๆจนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ ผู้วิจัยขอกราบขอบพระคุณและสำนึกในพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. ธีระพร วีระถาวร ในฐานะประธานกรรมการ อาจารย์ ดร.อรุณี กำลัง และ ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์ กรรมการสอบวิทยานิพนธ์ และ ผู้ช่วยศาสตราจารย์ ดร. บุญอ้อม โฉมที กรรมการภายนอกมหาวิทยาลัย ที่กรุณาตรวจแก้วิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น และขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติที่ให้โอกาสทางการศึกษา และประสิทธิประสาทความรู้ให้แก่ผู้วิจัยกระทั่งสำเร็จการศึกษา

ท้ายนี้ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งสนับสนุนในด้านการเงินและให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา รวมทั้งพี่สาว พี่ชาย ญาติๆ เพื่อนๆ ทุกคนที่ส่งเสริมและให้กำลังใจแก่ผู้วิจัยมาตลอด

สารบัญ

สารบัญตาราง

ตาราง หน้า

ตาราง หน้า

สารบัญภาพ

ภาพประกอบ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

โดยทั่วไปการศึกษาข้อมูลเกี่ยวกับเหตุการณ์สิ่งที่เราสนใจจะเกิดขึ้นหรือไม่ หรือเกิดขึ้นด้วยความน่าจะเป็นเท่าไร เช่นการเป็นโรคหัวใจกับการไม่เป็นโรคหัวใจ การตายหรือไม่ตายของแมลง การยิงปืนถูกเป้าหรือไม่ถูกเป้า การนั่งรถส่วนตัวไปทำงานหรือเลือกนั่งรถประจำทาง การศึกษาต่อหรือไม่ศึกษาต่อ ภาวการณ์เป็นหนี้กับการไม่เป็นหนี้ จะเห็นว่าเหตุการณ์ที่กล่าวมานั้น ล้วนเป็นเหตุการณ์ที่สามารถเกิดขึ้นได้ตลอดเวลา และตัวแปรตอบสนองของเหตุการณ์ตัวอย่างเหล่านั้น ทั้งหมดเป็นตัวแปรสุ่มแบบสองกลุ่ม นั่นก็คือ ตัวแปรตอบสนองเป็นตัวแปรเชิงกลุ่มที่มีสองลักษณะ คือ มีค่าเท่ากับ 1 คือ เหตุการณ์ที่เราสนใจ หรือ มีค่าเท่ากับ 0 คือ เหตุการณ์ที่เราไม่สนใจ ส่วนตัวแปรอธิบายมีทั้งตัวแปรต่อเนื่อง และไม่ต่อเนื่องก็ได้

จากปัญหาข้างต้นไม่สามารถใช้การวิเคราะห์ความถดถอยแบบปกติได้ เนื่องจากว่าการวิเคราะห์การถดถอยเป็นการวิเคราะห์ข้อมูลเพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรตอบสนอง (response variable) และตัวแปรอธิบาย(explanatory variable) ซึ่งการวิเคราะห์การถดถอยตัวแปรอธิบาย ต้องเป็นตัวแปรเชิงปริมาณ หรือ ตัวแปรเชิงกลุ่มก็ได้ ในขณะที่ตัวแปรตอบสนองต้องเป็นเชิงปริมาณ เพียงอย่างเดียว ถ้าเราทำการวิเคราะห์ด้วยความถดถอยแบบปกติก็จะทำให้ผลการวิเคราะห์ออกมาไม่น่าเชื่อถือ จากสถานการณ์ต่าง ๆ ดังที่ได้กล่าวไว้ข้างต้น จะเห็นได้ว่า ตัวแปรตอบสนองมี 2 กลุ่ม (binary response) ซึ่งในการวิจัยโดยส่วนใหญ่มีการใช้เทคนิคการวิเคราะห์ข้อมูลด้วยตัวแบบโลจิท เนื่องจากว่าตัวแบบโลจิทนั้น คำนวณได้ง่าย ซึ่งเป็นเหตุผลที่มีผู้ใช้ตัวแบบโลจิทเพิ่มขึ้น และมากกว่าการใช้ตัวแบบโพรบิท เช่น ปี 1990-1994 มีผลงานที่ใช้ตัวแบบโลจิท 311 ชิ้น ส่วนตัวแบบโพรบิทมีผลงานปรากฏ 127 ชิ้น (Cramer ,2003) และตัวแบบโลจิทสามารถตีความหมายในเทอมของ odds ได้ จึงเป็นเหตุให้นิยมการใช้ตัวแบบโลจิทมากกว่าตัวแบบอื่นๆ ดังนั้น ผู้วิจัยจึงได้นำ ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก มาใช้ในการวิเคราะห์ข้อมูล ซึ่งยังคงวัตถุประสงค์ และ แนวความคิดเหมือนกับการวิเคราะห์การถดถอยแบบปกติ นั่นก็คือ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนอง และ ตัวแปรอธิบาย เพื่อนำสมการที่ได้ไปประมาณค่าหรือพยากรณ์ค่าตัวแปรตอบสนอง เมื่อกำหนดค่าตัวแปรอธิบาย

เนื่องจากตัวแบบที่นำมาใช้แสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและ ตัวแปรอธิบายในการวิเคราะห์ของ ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ซึ่งในตัวแบบเชิงเส้นที่วางนัยทั่วไป (generalized linear models :GLM) เรียกว่า ลิงค์ฟังก์ชัน (link function) โดยมี g เป็นตัวเชื่อมระหว่างตัวแปรตอบสนอง Y_i และตัวแปรอธิบาย X_i ตัวอย่างงานทดลองเมื่อค่าตอบสนอง Y_i มีได้เพียง 2 ค่าคือ 0 และ 1 ดังนั้นสามารถเขียนได้เป็น

$$P(Y_i = 0) = 1 - p_i \qquad \text{และ} \qquad P(Y_i = 1) = p_i \tag{*}$$

สำหรับความน่าจะเป็นของตัวแปรไม่ตอบสนองและตัวแปรที่ตอบสนอง ตัวแบบเชิงเส้นมีความสำคัญในการนำไปประยุกต์ในงานทางด้านทฤษฎี สมมติ X สามารถอธิบาย Y ได้ จะปรากฏเส้นพยากรณ์เชิงเส้น η_i โดย

$$\eta_i = \beta_0 + \beta_1 X_i \qquad ,i = 1,2,\ldots,n$$

เมื่อยังไม่ทราบค่าสัมประสิทธิ์พารามิเตอร์ β_0, β_1 สำหรับตัวแปรสุ่มแบบ 2 กลุ่ม(binary random variables) ให้ฟังก์ชันของ g อยู่ในช่วง $[0,1]$ นำไปสู่ค่าจริง η ซึ่งอยู่ในช่วง $(-\infty, \infty)$

ดังนั้น

$$g(p_i) = \eta_i = \beta_0 + \beta_1 X_i, i = 1,2,\ldots,n$$

การเลือกลิงค์ฟังก์ชันสามารถหาได้โดยใช้ฟังก์ชันเชื่อม(link function) คือ

1. โลจิทหรือ โลจีสติคฟังก์ชัน

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

2. โพรบิทฟังก์ชัน

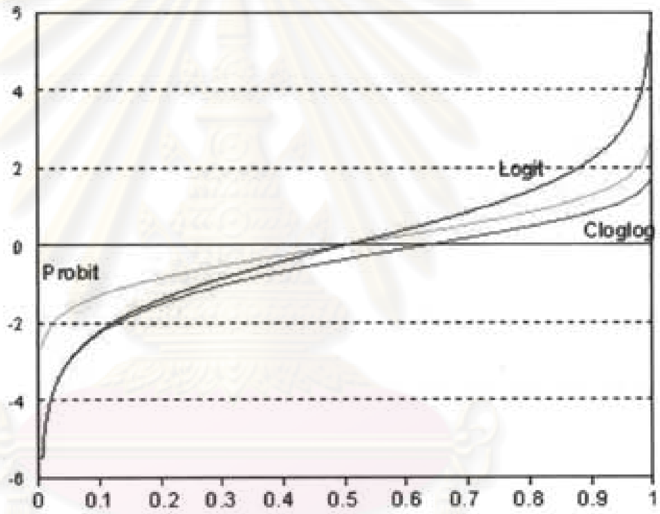
$$g(p) = \Phi^{-1}(p)$$

3. คอมพลีเมนทารีล็อก-ล็อก ฟังก์ชัน

$$g(p) = \log[-\log\{1-p\}]$$

ตารางที่ 1.1: ความน่าจะเป็นของ p_i เทียบกับฟังก์ชันค่าสังเกตของ x_i เมื่อลิงค์ฟังก์ชันต่างกัน

Link function	p_i	LD_p
logit	$\dfrac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$	$\dfrac{\log\left(\dfrac{p}{1-p}\right)-\beta_0}{\beta_1}$
probit	$\Phi(\beta_0+\beta_1 x_i)$	$\dfrac{\Phi^{1}(p)-\beta_0}{\beta_1}$
Complementary log-log	$1-\exp(-\exp(\beta_0+\beta_1 x_i))$	$\dfrac{\log(-\log(1-p))-\beta_0}{\beta_1}$



รูปที่ 1.1 : The three link functions by response probability

ที่มา : Seppo Laaksonen

ดังนั้น ถ้ามีความเข้าใจในเรื่องตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log model) เป็นอย่างดีแล้ว ก็จะสามารถวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ แต่เนื่องจาก β_0, β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า ถ้าสามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงก็ยิ่งทำให้การพยากรณ์ถูกต้องมากยิ่งขึ้น ผู้วิจัยจึงสนใจที่จะศึกษาถึงวิธีการประมาณค่าพารามิเตอร์ของตัวแบบ ด้วยวิธีการประมาณที่ใช้กันอย่างแพร่หลาย ได้แก่ วิธีภาวะน่าจะเป็นสูงสุด(maximum likelihood estimation method :MLE) วิธี

กำลังสองน้อยสุดถ่วงน้ำหนัก (weighted least squares method :WLS) และที่จะนำเสนออีกวิธีหนึ่ง นั้นก็คือ วิธีไคกำลังสองต่ำสุด(minimum chi-square method :MCS) ซึ่งเป็นงานวิจัยที่จะนำมาใช้ในประโยชน์ด้านต่างๆต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์ของตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log model) เมื่อตัวแปรตอบสนองเป็นแบบสองกลุ่ม โดยใช้วิธีการประมาณค่าพารามิเตอร์ 3 วิธี คือ

1. วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method : WLS)
2. วิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE)
3. วิธีไคกำลังสองต่ำสุด(Minimum Chi-Square method : MCS)

1.2.2 เพื่อเปรียบเทียบวิธีการสร้างตัวแบบโลจิท(Logit model) ตัวแบบโพรบิท(Probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(Complementary log-log model) เมื่อตัวแปรตอบสนองเป็นแบบสองกลุ่ม โดยใช้วิธีการประมาณค่าพารามิเตอร์ 3 วิธี คือ

1. วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method : WLS)
2. วิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE)
3. วิธีไคกำลังสองต่ำสุด(Minimum Chi-Square method : MCS)

โดยพิจารณาจากค่า Deviance ว่า Deviance ของตัวแบบไหนจากวิธีการประมาณค่า 3 วิธี ให้ค่า Deviance ต่ำสุดถือว่าวิธีการประมาณค่านั้นดีที่สุด

1.3 ขอบเขตการวิจัย

ในการวิจัยครั้งนี้ได้ศึกษาการวิเคราะห์การถดถอยเมื่อตัวแปรตอบสนองมีสองลักษณะ คือ 1 (เหตุการณ์ที่เราสนใจ) กับ 0 (เหตุการณ์ที่เราไม่สนใจ) โดยใช้ข้อมูลจริงจากหลากหลายสาขาวิชา เช่น ข้อมูลทางด้านการแพทย์ ด้านวิทยาศาสตร์ วิศวกรรมศาสตร์ และ ทางสังคมศาสตร์ เหตุผลที่เลือกข้อมูลทั้ง 9 ชุดนี้ เนื่องจากว่าข้อมูลโดยส่วนใหญ่เลือกใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล ทางผู้วิจัยจึงมีความสนใจที่จะนำข้อมูลทั้ง 9 ชุดมาวิเคราะห์ด้วย ตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log model) เพื่อทำการศึกษาวิธีการหาค่าประมาณ หรือค่าพยากรณ์ของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ 3 วิธี เมื่อกำหนดขอบเขตการวิจัยดังนี้

1. ข้อมูลชุดที่ 1 เป็นข้อมูลผู้อาศัยในหมู่บ้าน Amozonas ประเทศบราซิล ปี 1971 ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ เพื่อตรวจสอบในช่วงเวลาในการฉีดเซรุ่มป้องกันมาลาเรีย ว่า

เชรุ่มที่ให้ผลบวก หรือไม่ให้ผลบวก ข้อมูลจาก Draper,Voller and Carpenter(1972) ซึ่ง Draper ได้ใช้ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ในการวิเคราะห์ข้อมูล

2. ข้อมูลชุดที่ 2 เป็นข้อมูลของผู้สูบบุหรี่ที่ปราศจากสารกัมมันตภาพรังสีที่มีอายุระหว่าง 20-64 ปี ของบริษัท Coalminers ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ ว่ามีอาการหอบหรือไม่มีอาการหอบ เมื่อทำงานใน Coalminers ข้อมูลจาก Ashford and Sowden(1970)

3. ข้อมูลชุดที่ 3 เป็นข้อมูลผู้อาศัยเพศชายอายุ 40-59 ปี ในเมือง 2 เมือง ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความดันโลจิต เพื่อตรวจสอบในช่วง 6 ปี ต่อเนื่องกันว่าเป็นโรคหัวใจหรือไม่เป็นโรคหัวใจ ข้อมูลจาก Cornfield (1962) และ Agresti (1990) ซึ่งได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล

4. ข้อมูลชุดที่ 4 เป็นข้อมูลปริมาณความเข้มข้นของสารพิษที่มีผลต่อการตายของแมลง ข้อมูลจาก Martin(1942) และ Finney(1971)

5. ข้อมูลชุดที่ 5 เป็นข้อมูลทางชีววิทยาเกี่ยวกับการตายของแมลงเมื่อได้รับระดับความเข้มข้นที่ต่างกัน เป็นข้อมูลจริงของ Muhammad(1990) ซึ่ง Muhammad ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล

6. ข้อมูลชุดที่ 6 เป็นข้อมูลความเข้มข้นของแอมโมเนีย ที่มีผลต่อการตายของแมลง ข้อมูลจาก Strand(1930)

7. ข้อมูลชุดที่ 7 เป็นข้อมูลการเสียหายของหมุดเจาะบนเครื่องบิน เมื่อระดับความกดอากาศเพิ่มขึ้นที่ละ 200 psi จาก 2500-4300 psi ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความกดอากาศ ว่าความกดอากาศที่ระดับต่างจะมีผลต่อการเสียหายหรือไม่เสียหายของหมุดเจาะบนเครื่องบิน ข้อมูลจาก Montgomery and Peck(1982) ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล

8. ข้อมูลชุดที่ 8 เป็นข้อมูลการสำรวจทางสังคม ปี 1982 ของคนผิวขาว ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่อง Political Views เพื่อดูว่าคนผิวขาวจะเลือก Reagan เป็นประธานาธิบดี ข้อมูลจาก Clogg and Shockey (1988)

9. ข้อมูลชุดที่ 9 เป็นข้อมูลการสำรวจความคิดเห็นเกี่ยวกับบทบาทของผู้หญิงที่มีต่อสังคม ทำการสำรวจทั้ง เพศหญิงและเพศชาย ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องปีของการสำเร็จการศึกษาต่อความคิดเห็นของการเห็นด้วยหรือไม่เห็นด้วยของ คำกล่าวที่ว่า "ผู้หญิงมีหน้าที่ดูแลบ้านและอนุญาตให้ทำงานนอกบ้านได้เหมือนผู้ชาย" ข้อมูลจาก Haberman(1978) ได้ทำการตัดข้อมูลปีสำเร็จการศึกษาตั้งแต่ 0-2 เนื่องจากcell นี้มีค่าเป็นศูนย์ และ 19-20 ออกเพราะเนื่องจากความถี่ใน cell นี้มีค่ากระโดดไม่คงที่

10. ตัวแปรตอบสนองที่ใช้ในการศึกษา เป็นตัวแปรตอบสนองเชิงกลุ่มที่มีสองลักษณะ ซึ่งจำแนกกลุ่มของเหตุการณ์ออกเป็นสองกลุ่ม คือ เหตุการณ์ที่สนใจ และเหตุการณ์ที่ไม่สนใจ

11. ตัวแปรอธิบายที่ใช้ในการศึกษา คือ ตัวแปรแบบต่อเนื่องหรือไม่ต่อเนื่องก็ได้

12. ใช้ตัวสถิติ Deviance เป็นเกณฑ์ในการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์และเลือกตัวแบบ โดยที่

$$D = 2\sum_{i}\left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}$$

13. กำหนดระดับนัยสำคัญ 0.05

14. การประมวลผลใช้โปรแกรม R

1.4 เกณฑ์การตัดสินใจ

การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ว่าวิธีใดน่าจะมีความถูกต้องมากและตัวแบบใดเหมาะสมที่สุดจะพิจารณาจากเกณฑ์การเปรียบเทียบของ Deviance ดังนี้

$$D = 2\sum_{i}\left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}$$

โดยที่ y_i แทนค่าสังเกตที่ $i, i = 1,2,...,n$

$\hat{y}_i$ แทนค่าพยากรณ์ที่ $i, i = 1,2,...,n$

n_i แทนจำนวนตัวอย่างที่ใช้ในการทดลองในกลุ่มที่ $i, i = 1,2,...,n$

นำค่า Deviance ของทั้ง 3 วิธี มาเปรียบเทียบว่าวิธีการใดให้ค่า Deviance ของการประมาณค่าน้อยที่สุด จึงจะสรุปได้ว่าเป็นวิธีการประมาณค่าพารามิเตอร์ที่ดีที่สุดของแต่ละตัวแบบ

1.5.ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อให้สามารถสร้างตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log)เมื่อตัวแปรตอบสนองมีเพียง 2 กลุ่มและประมาณค่าพารามิเตอร์ด้วย วิธีกำลังสองน้อยแบบสุดถ่วงน้ำหนัก (Weighted Least Squares method : WLS)

2. เพื่อให้สามารถสร้างตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log)เมื่อตัวแปรตอบสนองมีเพียง 2 กลุ่มและประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE)

3. เพื่อให้สามารถสร้างตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log)เมื่อตัวแปรตอบสนองมีเพียง 2 กลุ่ม และประมาณค่าพารามิเตอร์ด้วย วิธีไคกำลังสองต่ำสุด (Minimum Chi-Square method : MCS)

4. เพื่อให้ทราบว่าวิธีการประมาณตัวแบบโลจิท(logit model) ตัวแบบโพรบิท(probit Model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก(complementary log-log)เมื่อตัวแปรตอบสนองมีเพียง 2 กลุ่ม ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบวิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE) วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method : WLS) หรือ วิธีไคกำลังสองต่ำสุด (Minimum Chi-Square method : MCS) 3 วิธีนี้วิธีไหนดีกว่ากัน ด้อยกว่ากันในกรณีใดบ้าง

5. เพื่อเป็นแนวทางในการนำไปประยุกต์เลือกใช้ตัวแบบให้มีความเหมาะสมกับลักษณะของข้อมูลเมื่อตัวแปรตอบสนองมีเพียง 2 กลุ่ม

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ตัวแบบที่ตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ (Binary Response Model)

ในการวิเคราะห์การถดถอยของตัวแบบที่ตัวแปรตอบสนอง (Y) มีค่าเป็น 2 ลักษณะ (binary response model) คือมีค่าเป็น 0 เมื่อไม่เกิดเหตุการณ์ที่สนใจ หรือมีค่าเป็น 1 เมื่อเกิดเหตุการณ์ที่สนใจ ส่วนตัวแปรอธิบาย (X) นั้นอาจเป็นได้ทั้งตัวแปรเชิงปริมาณ หรือตัวแปรเชิงคุณภาพ เรียกการวิเคราะห์การถดถอยของตัวแบบลักษณะนี้ว่า การวิเคราะห์การถดถอยทวิ (binary regression) ซึ่งจะมีวิธีการวิเคราะห์ที่แตกต่างไปจากการวิเคราะห์การถดถอยเชิงเส้นทั่วไป เนื่องจากข้อแตกต่างในลักษณะของค่าคาดหวัง (expected value) ของตัวแปรตอบสนอง และปัญหาที่ไม่เป็นไปตามข้อสมมติ (assumptions) ของการวิเคราะห์การถดถอยดังนี้

2.1.1 ค่าคาดหวังของตัวแปรตอบสนองกรณีมีตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ

เมื่อพิจารณาตัวแบบของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (simple linear regression) คือ

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad ; i = 1,2,\ldots,n \qquad (2.1)$$

โดยที่ Y_i = ค่าสังเกตที่ i ของตัวแปรตอบสนองซึ่งมีค่าได้ 2 ค่า คือ 0 และ 1

X_i = ค่าสังเกตที่ i ของตัวแปรอธิบาย

β_0, β_1 = ค่าสัมประสิทธิ์การถดถอย (regression coefficients)

ε_i = ค่าความคลาดเคลื่อนของค่าสังเกตที่ i

ถ้าข้อสมมติการถดถอยที่ว่า $E(\varepsilon_i) = 0$ เป็นจริงแล้ว จะได้ค่าคาดหวังของตัวแปรตอบสนอง คือ

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad (2.2)$$

ถ้าพิจารณา Y_i เป็นตัวแปรสุ่มแบบเบอร์นูลลี (Bernoulli) จะได้การแจกแจงความน่าจะเป็นดังนี้

Y_i	Probability
1	$Pr(Y_i = 1) = P_i$
0	$Pr(Y_i = 0) = 1 - P_i$

เมื่อ P_i คือความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

$1 - P$ คือความน่าจะเป็นของการเหตุการณ์ที่ไม่สนใจ

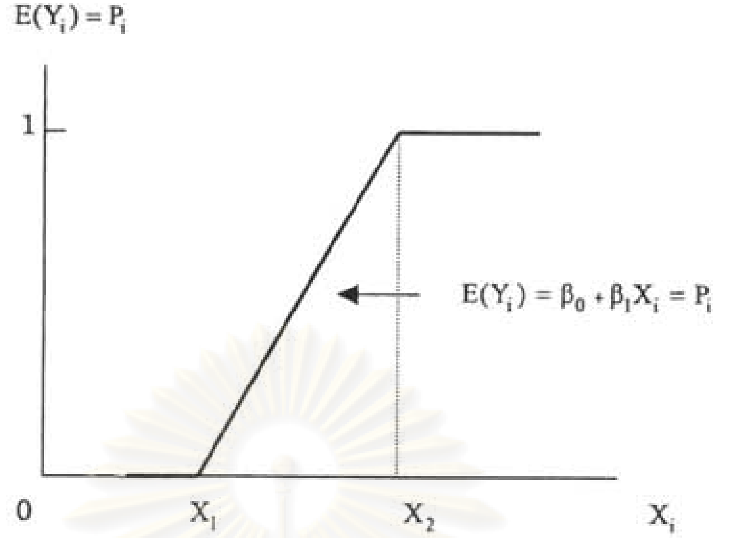
ดังนั้นจะได้ว่า

$$E(Y_i) = 1(P_i) + 0(1 - P_i) = P_i \tag{2.3}$$

จากสมการที่ (2.2)และสมการ (2.3) จะได้ค่าคาดหวังจากตัวแปรตอบสนองกรณีที่ตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะดังนี้

$$E(Y_i) = \beta_0 + \beta_1 X_i = P_i$$

นั่นแสดงว่าค่าคาดหวังของตัวแปรตอบสนองหรือ $E(Y_i)$เมื่อตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ ก็คือ " ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ ณ ที่ระดับของตัวแปรอธิบาย X_i " นั่นเอง ดังแสดงได้ในภาพที่ 2.1 ซึ่งสามารถอธิบายได้ว่า ค่าคาดหวังของตัวแปรตอบสนอง $E(Y_1)$ คือ ความน่าจะเป็นของเหตุการณ์ที่สนใจ ณ ที่ระดับของตัวแปรอธิบาย X_1 ซึ่งมีค่าเท่ากับ 0 และค่าคาดหวังของตัวแปรตอบสนอง $E(Y_2)$ คือ ความน่าจะเป็นของเหตุการณ์ที่สนใจ ณ ที่ระดับของตัวแปรอธิบาย X_2 ซึ่งมีค่าเท่ากับ 1



รูปที่ 2.1 แสดงความน่าจะเป็นที่ตัวแปรตอบสนองจะเกิดเหตุการณ์ที่สนใจ
เมื่อตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ
ที่มา : Neter et al (1989)

2.1.2 ปัญหาที่เกิดขึ้นเมื่อตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ

การวิเคราะห์การถดถอยมักทำตามตัวแบบและข้อสมมติของการวิเคราะห์การถดถอยเชิงเส้นทั่วไปที่มีตัวแปรตอบสนองเป็นค่าแบบต่อเนื่อง ในการวิเคราะห์การถดถอยดังกล่าวจะใช้การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด (ordinary least square : OLS) โดยมีข้อสมมติของการวิเคราะห์ดังนี้

ก. ข้อสมมติเกี่ยวกับค่าความคลาดเคลื่อน (ε_i)

1) (ε_i) มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 $[E(\varepsilon_i) = 0]$ และมีค่าความแปรปรวนเท่ากับ σ^2 $[Var(\varepsilon_i) = \sigma^2]$ นั่นคือ $\varepsilon_i \sim Normal(0, \sigma^2)$; $i = 1,2,\ldots,n$

2) ε_i และ ε_j มีการแจกแจงที่เป็นอิสระต่อกัน นั่นคือ $COV(\varepsilon_i, \varepsilon_j) = 0$ เมื่อ $i \neq j$

ข. ข้อสมมติเกี่ยวกับตัวแปรอธิบายคือ

1) ตัวแปรอธิบายแต่ละตัวจะต้องไม่มีความสัมพันธ์เชิงเส้นระหว่างกัน

2) ตัวแปรอธิบาย และค่าความคลาดเคลื่อน เป็นอิสระกัน

จากตัวแบบของการวิเคราะห์ของการถดถอยเชิงเส้นอย่างง่ายซึ่งมีรูปแบบดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad ; i = 1,2,\ldots,n$$

ถ้าข้อสมมติของการวิเคราะห์ถดถอยที่ว่า $[E(\varepsilon_i) = 0]$ เป็นจริงแล้ว จะทำให้ได้

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i + E(\varepsilon_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

โดยที่ Y_i เป็นผลลัพธ์ของค่าสังเกตจากการทดลองครั้งที่ i เมื่อ X มีค่าเท่ากับ X_i ดังนั้นฟังก์ชันการถดถอย (regression function) ของตัวแบบคือ

$$\begin{aligned} \mu_Y = E(Y) &= E(\beta_0 + \beta_1 X + \varepsilon) \\ &= \beta_0 + \beta_1 X \qquad ; \ E(\varepsilon) = 0 \end{aligned}$$

ซึ่งบอกถึงค่าเฉลี่ยของการแจกแจงความน่าจะเป็นของ Y ณ ค่าของ X ที่กำหนดให้ และความแปรปรวนคือ

$$\begin{aligned} Var(Y) &= Var(\beta_0 + \beta_1 X + \varepsilon) \\ &= Var(\varepsilon) = \sigma^2 \end{aligned}$$

จากข้อสมมติเกี่ยวกับค่าความคลาดเคลื่อนในข้อ ก) ที่ว่า ε มีการแจกแจงแบบปกติในตัวแบบการถดถอยข้างต้นทำให้หมายความได้ว่า Y มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยคือ $E(Y) = \beta_0 + \beta_1 X$ และความแปรปรวน คือ $Var(Y) = \sigma^2$ นั่นคือ

$$Y \sim Normal[E(Y), Var(Y)]$$

หรือ

$$Y \sim Normal[\beta_0 + \beta_1 X, \sigma^2]$$

และข้อสมมติที่ว่า ε_i และ ε_j เป็นอิสระต่อกัน นั่นคือ $COV(\varepsilon_i, \varepsilon_j) = 0$ เมื่อ $i \neq j$ เป็นจริงแล้ว จะทำให้ Y_i และ Y_j เป็นอิสระต่อกันด้วย นั่นคือ $COV(Y_i, Y_j) = 0$ เมื่อ $i \neq j$

ข้อสมมติของตัวแบบการถดถอยที่เกี่ยวข้องกับค่าความคลาดเคลื่อนมีความจำเป็นอย่างยิ่งต่อการทดสอบสมมติฐานและการประมาณค่าพารามิเตอร์ของตัวแบบ หากข้อสมมติข้อใดข้อหนึ่งไม่เป็นจริงจะมีผลทำให้ตัวประมาณที่ได้จากวิธีกำลังสองน้อยสุดไม่มีคุณสมบัติเป็นตัวประมาณที่ดี และการสรุปผลจากข้อสมมติฐานเกี่ยวกับข้อมูลที่นำมาวิเคราะห์จะผิดพลาดได้ ดังนั้นจึงควรมีการตรวจสอบข้อมูลที่นำมาศึกษาก่อนว่ามีคุณสมบัติตามข้อสมมติของตัวแบบหรือไม่

อย่างไรก็ตามถ้าตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะแล้ว จะทำให้เกิดปัญหาที่ไม่เป็นไปตามข้อสมมติของการวิเคราะห์การถดถอยเชิงเส้นทั่วไป ซึ่ง Neter et al (1989) ได้กล่าวถึงปัญหาที่เกี่ยวกับข้อสมมติของการวิเคราะห์การถดถอยเชิงเส้นเมื่อตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ ดังต่อไปนี้

2.1.2.1 ค่าความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติ (non-normal error terms)

เนื่องจาก ค่าความคลาดเคลื่อน (ε_i) เป็นฟังก์ชันของ Y_i นั่นคือ

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

ดังนั้นถ้า Y_i มีค่าเป็น 2 ลักษณะ คือ 0 และ 1 ซึ่ง Y_i มีการแจกแจงแบบเบอร์นูลลีแล้ว จะมีผลทำให้ค่าความคลาดเคลื่อนมีการแจกแจงแบบเบอร์นูลลีด้วย โดยมีค่าเป็น 2 ค่าเช่นกัน คือ

ถ้า $Y_i = 1 \rightarrow \quad \varepsilon_i = 1 - (\beta_0 + \beta_1 X_i)$

และถ้า $Y_i = 0 \rightarrow \quad \varepsilon_i = -(\beta_0 + \beta_1 X_i)$

เมื่อค่าความคลาดเคลื่อนมีการแจกแจงแบบไม่เป็นปกติจึงไม่เป็นไปตามข้อสมมติของการวิเคราะห์การถดถอยเชิงเส้นทั่วไป การแก้ปัญหาในส่วนนี้อาจทำได้โดยการใช้ตัวอย่างที่ใหญ่ ซึ่งจะทำให้ Y_i มีการแจกแจงโดยประมาณเป็นแบบปกติ และมีผลทำให้ค่าความคลาดเคลื่อนมีการแจกแจงโดยประมาณเป็นแบบปกติด้วย

2.1.2.2 ค่าความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ (non-constant error variance)

ปัญหาอีกอย่างหนึ่งของค่าความคลาดเคลื่อนคือ ความแปรปรวนของค่าความคลาดเคลื่อนจะมีค่าไม่คงที่ ถ้าตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ ถ้าพิจารณาจากตัวแบบของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายในสมการ (2.1) จะได้ความแปรปรวนของ Y_i คือ

$$
\begin{aligned}
\mathrm{Var}(Y_i) &= E\{Y_i - E(Y_i)\}^2 \\
&= (1-P_i)^2 P_i + (0-P_i)^2(1-P_i) \\
&= P_i(1-P_i) \\
&= E(Y_i)\{1-E(Y_i)\}
\end{aligned}
$$

และเนื่องจากความแปรปรวนของค่าความคลาดเคลื่อนมีค่าเดียวกับค่าความแปรปรวนของ Y_i ดังนั้นจะได้ว่า

$$
\begin{aligned}
\mathrm{Var}(\varepsilon_i) &= \mathrm{Var}(Y_i) \\
&= P_i(1-P_i) && (2.4) \\
&= E(Y_i)\{1-E(Y_i)\} \\
&= (\beta_0+\beta_1 X_i)(1-\beta_0-\beta_1 X_i) && (2.5)
\end{aligned}
$$

จากสมการที่ (2.4) เห็นได้ว่า $\mathrm{Var}(\varepsilon_i)$ แต่ละค่าแปรเปลี่ยนไปตามค่าของ P_i และจากสมการที่ (2.5) ค่าของ P_i แต่ละค่าแปรเปลี่ยนไปตามค่าของตัวแปรอธิบาย X_i ซึ่งทำให้ความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ จึงไม่เป็นไปตามข้อสมมติของการวิเคราะห์การถดถอยเชิงเส้นทั่วไป ดังนั้นการประมาณค่าพารามิเตอร์โดยใช้วิธีกำลังสองน้อยสุดนั้นจึงเป็นการไม่เหมาะสม

2.1.2.3 ขอบเขตของค่าคาดหวังของตัวแปรตอบสนอง

เนื่องจากค่าคาดหวังของตัวแปรตอบสนองในตัวแบบที่ตัวแปรตอบสนองมีค่าเป็น 2 ลักษณะ นั้นอยู่ในรูปแบบของความน่าจะเป็น (P) ซึ่งค่าความน่าจะเป็นนั้นจะมีค่าอยู่ระหว่าง 0 และ 1 จึงทำให้ค่าคาดหวังของตัวแปรตอบสนองมีขอบเขตอยู่ระหว่าง 0 และ 1 คือ

$$0 \le E(Y) = P \le 1$$

แต่ค่าคาดหวังของตัวแปรตอบสนองในตัวแบบของการถดถอยเชิงเส้นทั่วไปอาจมีค่าต่ำกว่า 0 หรือเกินกว่า 1 ได้ ซึ่งขึ้นอยู่กับตัวแปรอธิบายในตัวแบบนั้น ดังนั้นการวิเคราะห์การถดถอยทวิ โดยใช้ตัวแบบของการถดถอยเชิงเส้นทั่วไปจึงอาจทำให้เกิดปัญหาค่าคาดหวังของตัวแปรตอบสนองอยู่นอกช่วงของ 0 และ 1

ในการแก้ปัญหาต่างๆที่กล่าวถึงข้างต้นนั้น เราอาจใช้การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares : WLS) แทนวิธีกำลังสองน้อยสุด เพื่อแก้ปัญหาค่าความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ นอกจากนี้เราสามารถใช้ตัวอย่างขนาดใหญ่จะทำให้การประมาณพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดสามารถทำได้ และให้ค่าประมาณ

พารามิเตอร์มีการแจกแจงเข้าสู่แบบปกติเมื่อใกล้อนันต์ (asymptotically normal) ถึงแม้ว่าการแจกแจงของค่าความคลาดเคลื่อนจะไม่เป็นแบบปกติก็ตาม และสำหรับการแก้ปัญหาที่ค่าคาดหวังของตัวแปรตอบสนองอยู่นอกช่วงของ 0 และ 1 นั้น สามารถทำได้ด้วยการวิเคราะห์โดยใช้ตัวแบบการถดถอยความน่าจะเป็น (probability regression model) ซึ่งได้แก่ ตัวแบบการถดถอยโลจิสติค (logistic regression model) หรืออาจใช้ ตัวแบบถดถอยโพรบิท (probit regression model) ซึ่งจะทำให้ค่าคาดหวังของตัวแปรตอบสนองมีค่าจำกัดอยู่ภายในช่วงของ 0 และ 1 (พิมลรัตน์ ; 2547) และอีกทางเลือกหนึ่งคือตัวแบบคอมพลีเมนทารี ล็อก-ล็อก (complementary log-log model)

2.2. ตัวแบบโลจิท

ตัวแบบโลจิทถูกใช้อย่างกว้างขวางทางด้านสังคมศาสตร์และชีววิทยา ตัวแบบนี้ใช้ประโยชน์ในกรณีเฉพาะด้านการวิจัยระบาดวิทยา(epidemiological) และประชากรศาสตร์ (demography) ในการประเมินผลกระทบของปัจจัยอธิบายบนความเสี่ยงสัมพัทธ์(relative risk) ของอัตราการเกิด อัตราการตาย และระยะเริ่มต้นของโรคหรือการเจ็บป่วย การแปลงโลจีสติค (logistic transformation) สามารถอธิบายด้วย logarithm ของ odds ของความสำเร็จ กับ ความไม่สำเร็จ การแปลงโลจีสติค (logistic transformation) ของความน่าจะเป็นของความสำเร็จ p คือ

$$\mathrm{logit}(p_i) \quad = \quad \log\left(\frac{p_i}{1-p_i}\right) \qquad (2.6)$$

สมการ(2.6) ลิงค์ฟังก์ชัน (link function) ใน ตัวแบบเชิงเส้นที่วางนัยทั่วไป (generalized linear models :GLM) และค่าที่หาได้จากตัวแบบโลจิท คือ

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i \; = \; \beta_0 + \beta_1 x_i \qquad (2.7)$$

สามารถอธิบายความน่าจะเป็น p_i ได้ดังนี้

$$p_i \quad = \quad \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} \; = \; \Lambda(\eta_i) \qquad (2.8)$$

เมื่อ $\Lambda(\eta_i)$ แทนฟังก์ชัน $\dfrac{\exp(\eta_i)}{1+\exp(\eta_i)}$

สำหรับค่าที่เป็นบวกของ x และ β การแปลงโลจิสติค (logistic transformation) ทำให้แน่ใจว่าค่า p อยู่ในช่วง $[0,1]$, $p \to 0$ $\text{logit}(p)$ มีแนวโน้มเข้าสู่ $-\infty$ และ $p \to 1$ $\text{logit}(p)$ มีแนวโน้มเข้าสู่ $+\infty$ (Powers ,2000)

2.2.1 ตัวแบบถดถอยโลจิสติค (Logistic Regression Model)

กำหนดให้ $P(x) = E(Y|X)$ คือ ค่าเฉลี่ยแบบมีเงื่อนไขของตัวแปรตอบสนอง Y ซึ่งเป็นตัวแปรเชิงกลุ่ม เช่น 2 กลุ่ม (dichotomous) หรือหลายกลุ่ม (polytomous) เมื่อกำหนดตัวแปรอธิบาย X มาให้ (conditional mean of y given x) การแจกแจงของตัวแปรตอบสนอง (Y) เป็นแบบทวินาม (binomial distribution) หรือ $\text{Bin}(1, P(x))$ เมื่อพิจารณาที่ยังไม่ได้จัดกลุ่ม (ungrouped data) และ เป็น $\text{Bin}(n_i, P_i(x))$ เมื่อพิจารณาข้อมูลแบบจัดกลุ่ม (grouped data) โดย $P(x)$ แทนความน่าจะเป็นของตัวแปรตอบสนองในกลุ่มที่สนใจศึกษา ในกรณีที่ตัวแปรตอบสนองเป็นตัวแปรเชิงกลุ่ม 2 ระดับ คือ $Y = 0$ เป็นเหตุการณ์ที่ไม่สนใจ และ $Y = 1$ เป็นเหตุการณ์ที่สนใจ เมื่อ $Y = 1$ ความน่าจะเป็นของตัวแปรตอบสนองคือ $P(x)$ และเมื่อ $Y = 0$ ความน่าจะเป็นคือ $1 - P(x)$ ถ้าตัวแปรอธิบาย X มีเพียง 1 ตัวจะได้ตัวแบบถดถอยโลจิสติค

$$P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

และเพื่อที่จะให้ค่าของ $P(x)$ หรือ $E(Y|X)$ ที่มีค่าอยู่ระหว่าง 0 ถึง 1 และค่าของ x สามารถมีได้ไม่จำกัด จึงมีความจำเป็นจะต้องแปลง $P(x)$ ให้อยู่ในรูปแบบอื่นที่เข้าใจง่าย และมีคุณสมบัติตามต้องการ การแปลงที่ว่านี้ใช้ **logit transformation** หรือ $g(x)$

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

ตัวแบบคือ **ลอกการิทึมของ odds** (อัตราส่วนความน่าจะเป็นของสำเร็จต่อความน่าจะเป็นไม่สำเร็จ) ในรูปความสัมพันธ์เชิงเส้นกับตัวแปรอธิบาย เรียกตัวแบบนี้ว่า **ตัวแบบโลจิท** (logit model)

ดังนั้น

$$\text{odds} = \left[\frac{P(x)}{1 - P(x)}\right] = e^{\beta_0 + \beta_1 x}$$

ตัวแบบข้างต้นนี้ **มีรูปแบบเช่นเดียวกันกับ odds ของตัวแปรตอบสนองเมื่อ** $Y = 1$ **ด้วย**

$$P(x) = [1 - P(x)]e^{\beta_0 + \beta_1 x}$$

$$e^{\beta_0 + \beta_1 X} = P(x)[1 + e^{\beta_0 + \beta_1 x}]$$

นั่นคือ
$$P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

หรือ
$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

เรียกตัวแบบนี้ว่า **ฟังก์ชันการถดถอยโลจิสติค** (logistic regression function) หรือ **ตัวแบบโลจิสติค** (logistic or logistic regression model)

สำหรับการแจกแจงของ Y เมื่อพิจารณาแต่ละหน่วยของ Y ที่มีผลลัพธ์เป็น 0 หรือ 1 กรณีนี้ Y คือ**ตัวแปรเชิงสุ่มเบอร์นูลลี** (Bernoulli random variable) **ที่มีค่าเฉลี่ย** E(Y) โดยที่

$$
\begin{aligned}
E(Y) &= \sum YP(Y) \\
E(Y) &= 1 \times P(Y=1) + 0 \times P(Y=0) \\
&= P(Y=1|X=x) \\
&= P(x)
\end{aligned}
$$

หรือ $E(Y|X) = P(x)$ โดยมีค่าขึ้นอยู่กับ X

ดังนั้น
$$
\begin{aligned}
E(Y^2) &= \sum Y^2 P(y) \\
E(Y^2) &= 1^2[P(x)] + 0^2[1 - P(x)] \\
&= P(x)
\end{aligned}
$$

ฉะนั้นความแปรปรวนของ Y คือ Var(Y) โดยที่

$$
\begin{aligned}
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= P(x)[1 - P(x)]
\end{aligned}
$$

นอกจากนี้ เมื่อ $Y_i \overset{d}{\sim} bin[n_i, P_i(x)]$ โดยที่ $i = 1,2,\ldots,N$

$$
\begin{aligned}
E(Y_i) &= n_i P_i(x) \\
Var(Y) &= n_i P_i(x)[1 - P_i(x)]
\end{aligned}
$$

จะเห็นได้ว่า $g(x)$ หรือ logit มีรูปแบบเหมือนกับการถดถอยเชิงเส้นตรง (linear regression) ฉะนั้นในการคำนวณใดๆ เกี่ยวกับการถดถอยโลจีสติคหรือตัวแบบโลจิท จึงสามารถใช้คุณสมบัติของการถดถอยเชิงเส้นตรงได้ กล่าวคือ การวิเคราะห์การถดถอยโลจีสติค สามารถใช้วิธีการวิเคราะห์การถดถอยเชิงเส้นแบบง่ายหรือแบบพหุคูณ

2.3 ตัวแบบโพรบิท

ตัวแบบโพรบิทจัดเป็นทางเลือกหนึ่งของตัวแบบโลจิท ตัวแบบไม่เป็นเชิงเส้น (nonlinear model) ใน p คือการแปลง (transformation) ดังนั้น ฟังก์ชัน monotonic (monotonic function) ของ p เป็นเชิงเส้นที่มีความสัมพันธ์กับตัวแปรอธิบาย ดังนั้นความน่าจะเป็นของแถวที่ i หรือค่าสังเกตที่ i p_i เป็นฟังก์ชันการแจกแจงสะสมแบบปกติมาตรฐาน (standard cumulative normal distribution function)

$$\mathrm{p_i} = \int_{-\infty}^{\eta_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu^2\right) \mathrm{du} \qquad (2.9)$$

สมการที่(2.9) สามารถเปรียบเทียบได้กับสมการ (2.7) ของตัวแบบโลจิท และสมการ(2.9) สามารถเขียนใหม่ในรูปของ $p_i = \Phi(\eta_i)$ เมื่อ $\Phi(.)$คือ ฟังก์ชันการแจกแจงสะสมของการแจกแจงปกติมาตรฐาน (cumulative distribution function of the standard normal distribution) การแปลงโพรบิท หรือ นอร์มิท (probit (or normit) transformation) หรือ probit link ให้ส่วนผกผันของฟังก์ชันการแจกแจงสะสมแบบปกติมาตรฐาน (standard cumulative normal distribution function) สามารถอธิบายสมการ(2.9) ของ (η_i) คือ

$$\eta_i = \Phi^{-1}(\mathrm{p_i}) = \mathrm{probit}(\mathrm{p_i}) \qquad (2.10)$$

สมการ(2.11) นิยามถึงโพรบิทลิงค์ (probit link) ดังนั้นตัวแบบโพรบิทสามารถเขียนใหม่เป็น

$$\Phi^{-1}(\mathrm{p}_i) = \eta_i = \beta_0 + \beta_1 \mathrm{x_i} \qquad (2.11)$$

หรือ

$$p_i = \Phi(\beta_0 + \beta_1 \mathrm{x_i}) \qquad (2.12)$$

ลิงค์ของโลจิสติคฟังก์ชันและโพรบิทฟังก์ชันสมมาตรรอบ $p = 0.5$ เมื่อ $\text{logit}(p)$ และ $\text{probit}(p)$ เป็นศูนย์ทั้งคู่ $p \to 0$ $\text{probit}(p)$มีแนวโน้มเข้าสู่ $-\infty$ และ $p \to 1$ $\text{probit}(p)$ มีแนวโน้มเข้าสู่ $+\infty$ (Powers ,2000)

พิจารณาตัวอย่างเกี่ยวกับการทดลองประสิทธิภาพของยาฆ่าแมลงต่อไปนี้

ให้ x แทนปริมาณของสารเคมีที่เป็นสารพิษ (อาจใช้ค่าของ x หรือ $\log x$) และให้ D แทนปริมาณของสารเคมีน้อยสุดที่ทำให้สิ่งมีชีวิตตาย โดยที่ $Y = 1$ จะสมมูลกับ $(x \geq D)$ เช่น แมลงผักจะไม่ตาย ถ้าฉีดสารเคมีฆ่าแมลงในปริมาณของ $(x < D)$ และจะตายถ้า $(x \geq D)$ โดยค่าของ D อาจเปลี่ยนแปลงไปตามหน่วยทดลอง (วีรานันท์, 2544 : 87-89)

ให้ $G(d) = \Pr(D \leq d)$ แทน ฟังก์ชันการแจกแจงสะสมแบบปกติมาตรฐาน (standard normal cumulative distribution function) สำหรับการแจกแจงของประชากร ดังนั้น ณ ปริมาณคงที่ x ความน่าจะเป็นที่หน่วยทดลองหนึ่งๆที่ได้รับสารเคมีแล้วตาย คือ

$$\begin{aligned}\Pr(Y = 1) &= P(x)\\ &= \Pr(D \leq x)\\ &= G(x)\end{aligned}$$

ถ้า F คือ cdf ของการแปลงเชิงเส้นของ D เช่น ฟังก์ชันการแจกแจงสะสมแบบปกติมาตรฐาน (standard normal cumulative distribution function) ของกลุ่ม (family) ที่มี G เป็นสมาชิกดังนั้น ความน่าจะเป็นข้างต้นจะอยู่ในรูปแบบของ $F(\beta_0 + \beta_1 x)$ ซึ่งสามารถอธิบายได้ดังนี้

พิจารณาตัวอย่างการทดลองที่เกี่ยวกับยาฆ่าแมลงหรือสารพิษ(x) การแจกแจงของ log(x) จะใกล้เคียงการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 ถ้าให้ G แทน cdf ของการแจกแจงแบบปกติ จะพบว่า

$$\begin{aligned}P(x) &= G(x)\\ &= \Phi[(x-\mu)/\sigma]\end{aligned}$$

โดยที่ Φ แทน ฟังก์ชันการแจกแจงสะสมแบบปกติมาตรฐาน (standard normal cumulative distribution function) ซึ่งสามารถแปลงให้อยู่ในเทอมของ $P(x) = F(\beta_0 + \beta_1 x)$ โดยที่ $F = \Phi$, $\beta_0 = -\mu/\sigma$ และ $\beta_1 = 1/\sigma$ การแปลงตัวแบบข้างต้นจะมีผลให้ตัวแบบมีรูปแบบใหม่ ซึ่งเรียกว่า **ตัวแบบโพรบิท** (probit model)

$$\Phi^{-1}[P(x)] = \beta_0 + \beta_1 x \tag{2.13}$$

สำหรับตัวแบบโพรบิท เส้นโค้งของ P(x) (หรือ 1 - P(x) เมื่อ $\beta_1 < 0$) จะมีรูปลักษณะของ cdf แบบปกติด้วยค่าเฉลี่ย $\mu = -\beta_0 / \beta_1$ และส่วนเบี่ยงเบนมาตรฐาน $\sigma = 1/|\beta_1|$ และเนื่องจาก 68% ของพื้นที่โค้งปกติคือบริเวณพื้นที่ ภายในช่วง 1 เท่าของ σ จากค่าเฉลี่ย ดังนั้น $1/|\beta_1|$ จึงเป็นระยะระหว่างค่า x ทั้งหลาย โดยที่ P(x) = 0.6 หรือ 0.84 และ P(x) = 1/2 นอกจากนี้ Agresti (1990) กล่าวว่า อัตราการเปลี่ยนแปลงของ P(x) ณ ค่า x หนึ่งๆ คือ

$$\frac{\partial P(x)}{\partial x} = \beta_1 \phi(\beta_0 + \beta_1 x)$$

โดยที่ ϕ แทน **ฟังก์ชันความหนาแน่นแบบปกติมาตรฐาน** (standard normal density function) และอัตราส่วนดังกล่าวมีค่าสูงสุดเมื่อ

$$\beta_0 + \beta_1 x = 0$$

หรือ เมื่อ $x = -\beta_0 / \beta_1$ โดยอัตราที่สูงสุดมีค่าเท่ากับ

$$\beta_1 /(2\pi)^{1/2} \quad \text{หรือ} \quad 0.40\beta_1$$

นั้นคือ ณ จุดที่ P(x) = 1/2

การเปรียบเทียบตัวแบบโพรบิทกับตัวแบบโลจิสติค พบว่า อัตราการเปลี่ยนแปลงของ P(x) เมื่อ P(x) = 1/2 จะเท่ากันสำหรับ cdf ของทั้งเส้นโค้งโพรบิทและเส้นโค้งโลจิสติค เมื่อ β_1 ของทั้ง 2 ตัวแบบมีค่าดังนี้

$$\beta_1(\text{ของโลจิสติค}) = 0.40/0.25 = 1.6 \text{ เท่าของ } \beta_1(\text{ของโพรบิท})$$

และส่วนเบี่ยงเบนมาตรฐานจะเท่ากันเมื่อ

$$\beta_1(\text{ของโลจิสติค}) = \pi / \sqrt{3} = 1.8 \text{ เท่าของ } \beta_1(\text{ของโพรบิท})$$

เมื่อตัวแบบทั้งสองสามารถอธิบายข้อมูลได้ดี (fit well) ตัวประมาณค่าพารามิเตอร์ของตัวแบบโลจิสติคจะมีค่าประมาณ 1.6-1.8 เท่าของตัวประมาณพารามิเตอร์ของตัวแบบโพรบิท

2.4 ตัวแบบคอมพลีเมนทารีล็อก-ล็อก

ตัวแบบคอมพลีเมนทารีล็อก-ล็อก (complementary log-log model) เป็นส่วนขยายจากตัวแบบโลจิทและตัวแบบโพรบิท เมื่อค่าของ P(x) **เพิ่มจาก 0 ค่อนข้างช้าแต่มีค่าเข้าใกล้ 1 อย่างรวดเร็ว** (วีรานันท์ 2544 ; 89-93)

สำหรับฟังก์ชัน **link** ที่ใช้สำหรับตัวแบบโลจิทและตัวแบบโพรบิท จะมีคุณสมบัติ**สมมาตร**(symmetric) รอบค่า 0.5 หรือ

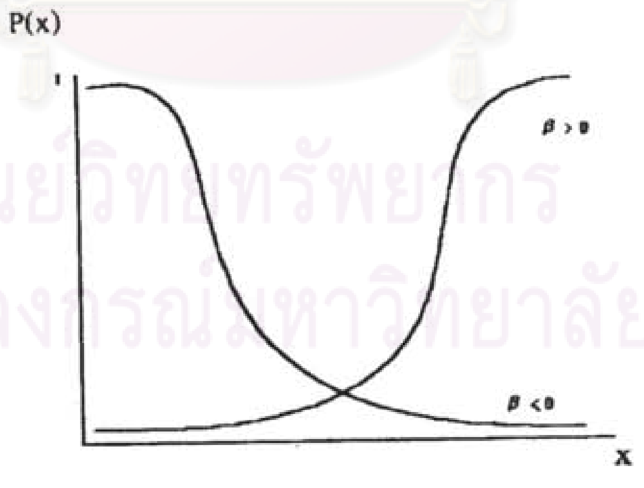
$$\text{link}[P(x)] = -\text{link}[1 - P(x)]$$

นั่นคือ

$$\begin{aligned}\text{logit}[P(x)] &= \log[P(x)/\{1 - P(x)\}] \\ &= \log P(x) - \log[1 - P(x)] \\ &= -\log[\{1 - P(x)\}/P(x)] \\ &= -\text{logit}[1 - P(x)]\end{aligned}$$

หมายความว่า โค้งของ P(x) สำหรับตัวแบบโลจิทและตัวแบบโพรบิท มีรูปแบบสมมาตรรอบจุด P(x) = 0.5 โดยเฉพาะ P(x) จะมีค่าเข้าใกล้ 0 ด้วย อัตราที่เท่ากันกับ เมื่อ P(x) มีค่าเข้าใกล้ 1

แต่ถ้าค่าของ P(x) **เพิ่มขึ้นจาก 0 ค่อนข้างช้า แต่มีค่าเข้าใกล้ 1 อย่างรวดเร็ว**(รูปที่ 2.2) ด้วยตัวแบบโลจิทและตัวแบบโพรบิท จะไม่เหมาะกับข้อมูล จึงควรใช้ตัวแบบอย่างอื่นคือตัวแบบคอมพลีเมนทารีล็อก-ล็อก และลักษณะของกราฟของ P(x) สำหรับตัวแบบคอมพลีเมนทารีล็อก-ล็อกแสดงไว้ในรูปที่ 2.2



รูปที่ 2.2 กราฟของ P(x) สำหรับตัวแบบคอมพลีเมนทารีล็อก-ล็อก

ลักษณะกราฟของ P(x) ข้างต้นนี้ ควรใช้เส้นโค้งของฟังก์ชัน(2.14) คือ

$$P(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)] \qquad (2.14)$$
$$1 - P(x) = \exp[-\exp(\beta_0 + \beta_1 x)]$$

ซึ่งมีรูปแบบไม่สมมาตรคือ P(x) มีค่าลดจาก 1 รวดเร็วกว่าการเข้าใกล้ 0 โดยสอดคล้องกับรูป 2.2 ฟังก์ชัน(2.14) นำไปสู่ **ตัวแบบคอมพลีเมนทารีล็อก-ล็อก** ใน (2.15) คือ

$$-\log[1 - P(x)] = \exp(\beta_0 + \beta_1 x)$$
$$\log[-\log\{1 - P(x)\}] = \beta_0 + \beta_1 x \qquad (2.15)$$

อนึ่งในทุกสมการ อาจใช้ log ฐาน 10 แต่โดยทั่วไปใช้ log ฐาน e ถึงแม้จะเขียน log ก็ตาม

การตีความหมายในตัวแบบคอมพลีเมนทารีล็อก-ล็อก สำหรับ x_1 และ x_2 ใดจะพบว่า

$$\log[-\log\{1 - P(x_2)\}] - \log[-\log\{1 - P(x_1)\}] = \beta_1(x_2 - x_1)$$

และ
$$\frac{\log[1 - P(x_2)]}{\log[1 - P(x_1)]} = \exp[\beta_1(x_2 - x_1)]$$

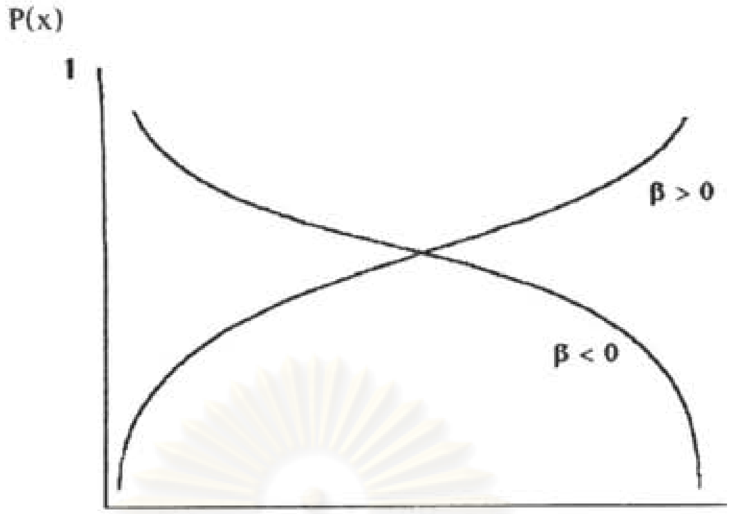
ดังนั้น
$$1 - P(x_2) = [1 - P(x_1)]^{\exp[\beta_1(x_2 - x_1)]}$$

ซึ่งหมายความว่า ความน่าจะเป็นที่ไม่สำเร็จ (probability of failure) ณ x_2 มีค่าเท่ากับความน่าจะเป็นที่ไม่สำเร็จ ณ x_1 ยกกำลัง $\exp[\beta_1(x_2 - x_1)]$

นอกจากตัวแบบ (2.14) หรือ (2.15) แล้ว ยังมีตัวแบบที่น่าสนใจและเกี่ยวกับหัวข้อนี้คือ ตัวแบบที่อยู่ในรูปแบบดังนี้

$$P(x) = \exp[-\exp(\beta_0 + \beta_1 x)]$$
$$\log[P(x)] = -\exp(\beta_0 + \beta_1 x)$$
$$-\log[P(x)] = \exp(\beta_0 + \beta_1 x)$$
$$\log\{-\log[P(x)]\} = \beta_0 + \beta_1 x \qquad (2.16)$$

ฟังก์ชัน (2.16) มีลักษณะของกราฟของ P(x) ที่มีค่าของ P(x) จะลดจากหนึ่งค่อนข้างช้า แต่จะเข้าใกล้ 0 อย่างรวดเร็ว ดังเส้นโค้งของรูปที่ 2.3 ดังนี้



รูปที่ 2.3 กราฟของ P(x) สำหรับตัวแบบล็อก-ล็อก

โดยค่าของ P(x) จะลดจาก 1 ค่อนข้างช้า แต่เข้าใกล้ 0 อย่างรวดเร็ว

ดังนั้นกรณีที่ใช้ link แบบล็อก-ล็อก (log-log link) จะนำไปสู่ **ตัวแบบล็อก-ล็อก** (log-log model) ซึ่งมีตัวแบบดัง (2.16) และรูปที่ 2.3

สรุปว่า ถ้าเป็น **ตัวแบบคอมพลีเมนทารีล็อก-ล็อก** จะใช้สำหรับความน่าจะเป็นที่สำเร็จ (probability of success) ส่วน **ตัวแบบล็อก-ล็อก** จะใช้สำหรับความน่าจะเป็น**ที่ไม่สำเร็จ**

หมายเหตุ : ตัวแบบล็อก-ล็อก $P(x) = \exp[-\exp(\beta_0 + \beta_1 x)]$ เป็น**กรณีพิเศษ**ของตัวแบบ $P(x) = F(\beta_0 + \beta_1 x)$ ด้วย cdf ของการแจกแจงค่าสุดโต่ง [extreme value (or Gumbel) distribution] เท่ากับ G(x) โดยที่

$$G(x) = \exp[-\exp\{-(x - a)/b\}]$$

ซึ่งมีค่าเฉลี่ยเท่า $a + 0.577b$ และส่วนเบี่ยงเบนมาตรฐานเท่ากับ $\pi b/\sqrt{6}$ สำหรับพารามิเตอร์ $b > 0, -\infty < a < +\infty$ นอกจากนี้ตัวแบบล็อก-ล็อก ยังสามารถประมาณได้ด้วย Fisher scoring algorithm for GLMs เช่นเดียวกัน

อนึ่งตัวแบบโลจิทและตัวแบบโพรบิท จะสมมาตรรอบ $P(x) = 0.5$ สามารถตรวจสอบได้ดังนี้

$$
\begin{aligned}
\because \quad \operatorname{logit}P(x) &= \ln\left(\frac{P(x)}{1-P(x)}\right) \\
&= \ln P(x) - \ln(1-P(x)) \\
&= -\ln\left(\frac{1-P(x)}{P(x)}\right) \\
\therefore \quad \operatorname{logit}P(x) &= -\operatorname{logit}(1-P(x))
\end{aligned}
$$

ส่วนตัวแบบคอมพลีเมนทารีล็อก-ล็อก และตัวแบบล็อก-ล็อก จะไม่สมสาตรรอบ $P(x) = 0.5$ เช่น ถ้า $P(x)$ มีค่าเพิ่มจาก 0 ช้าๆ แต่มีค่าเข้าใกล้ 1 อย่างรวดเร็ว ตัวแบบจะอยู่ในรูปของสมการ (2.14) คือ

$$P(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)]$$

หรือ

$$1 - P(x) = \exp[-\exp(\beta_0 + \beta_1 x)]$$

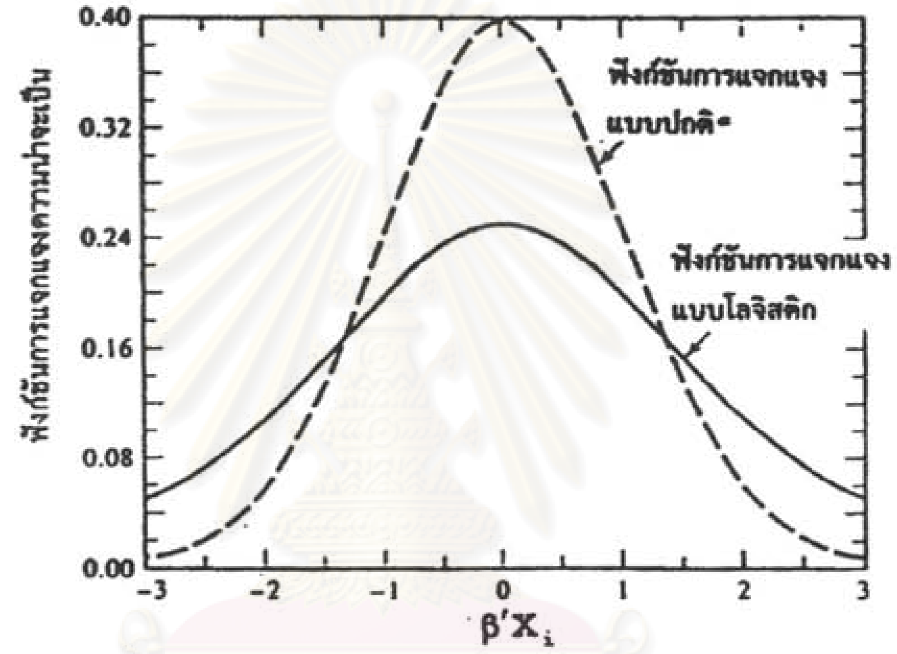
กราฟสำหรับ $P(x)$ จะมีลักษณะดังรูป 2.2 สำหรับ $1 - P(x) = Q(x)$ จะหมายความว่า $Q(x)$ มีค่าเพิ่มจาก 0 อย่างรวดเร็วแต่มีค่าใกล้ 1 อย่างช้าๆ ซึ่งมีลักษณะของกราฟดังรูปที่ 2.3

สรุปจากข้างต้น คือ $P(x)$ หรือ (2.14) คือตัวแบบคอมพลีเมนทารีล็อก-ล็อก ส่วน $1 - P(x)$ หรือ (2.16) ตัวแบบล็อก-ล็อก

2.5 ข้อเปรียบเทียบตัวแบบโลจิสติค ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก

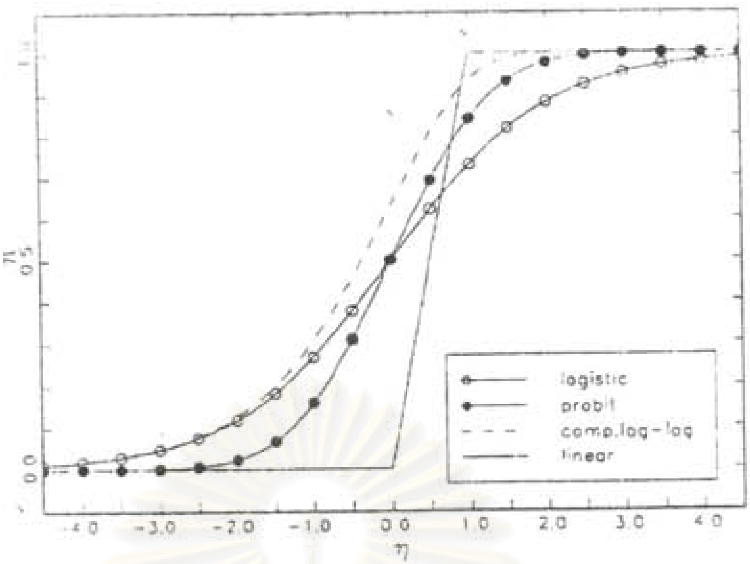
การหาความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอธิบายเมื่อตัวแปรตอบสนองเป็นตัวแปรเชิงกลุ่มที่มีสองค่า ส่วนตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณ สามารถใช้การวิเคราะห์ที่มีลักษณะไม่เป็นเชิงเส้นของตัวแบบโพรบิทและตัวแบบคอมพลีเมนทารีล็อก-ล็อก ที่มีอยู่บนรากฐานของการแจกแจงสะสมปกติมาตรฐาน หรือตัวแบบโลจิทที่อยู่บนรากฐานของการแจกแจงโลจิสติคได้ ซึ่งตัวแบบทั้งสามมีคุณสมบัติคล้ายกัน การเลือกใช้ตัวแบบใดขึ้นกับพื้นฐานทางทฤษฎีซึ่งจะมีข้อแตกต่างดังนี้

1. ลักษณะของฟังก์ชันการแจกแจงความน่าจะเป็น (รูปที่ 2.4) กราฟฟังก์ชันการแจกแจงความน่าจะเป็นแบบปกติมาตรฐาน และกราฟการแจกแจงความน่าจะเป็นแบบโลจิสติค จะมีลักษณะเป็นรูประฆังคว่ำโดยสมมาตรที่ค่าเฉลี่ยตัวแปรตอบสนอง (y_i) เท่ากับ 0 มีค่ามัธยฐาน ค่าฐานนิยม และค่าความเบ้เท่ากับ 0 โดยฟังก์ชันการแจกแจงความน่าจะเป็นแบบปกติมาตรฐานมีค่าเฉลี่ยเท่ากับ 0 ค่าความแปรปรวนเท่ากับ 1 ส่วนฟังก์ชันการแจกแจงความน่าจะเป็นแบบโลจีสติคมีค่าเฉลี่ยเท่ากับ 0 ค่าความแปรปรวนเท่ากับ $\frac{\pi^2}{3}$ ในขณะที่ฟังก์ชันการแจกแจงความน่าจะเป็นแบบคอมพลีเมนทารีล็อก-ล็อกมีค่าเฉลี่ยเท่ากับ -0.5772 ค่าความแปรปรวนเท่ากับ $\frac{\pi^2}{6}$



รูปที่ 2.4 กราฟฟังก์ชันการแจกแจงความน่าจะเป็นแบบปกติมาตรฐาน
เปรียบเทียบกับฟังก์ชันการแจกแจงความน่าจะเป็นแบบโลจีสติค
ที่มา: Griffiths และคณะ(1993)

2. ลักษณะฟังก์ชันการแจกแจงสะสม (รูปที่ 2.5) กราฟจะมีลักษณะในรูปตัวเอส ทั้งตัวแบบโพรบิทและตัวแบบโลจีสติคมีค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจและความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจใกล้กันมากจนเท่ากัน ฉะนั้นการใช้ตัวแบบใดก็จะไม่ก่อให้เกิดความแตกต่างในการหาค่าความน่าจะเป็น ยกเว้นแต่กรณีที่มีข้อมูลหนาแน่นในช่วงหาง (ชนิวศรา ; 2543) ในขณะที่ฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก มีความชันสูงกว่าฟังก์ชันโลจีสติคและฟังก์ชันโพรบิทที่ปรับแล้ว และเมื่อค่าของ P(x) ของตัวแบบคอมพลีเมนทารีล็อก-ล็อก เพิ่มจาก 0 ค่อนข้างช้าแต่มีค่าเข้าใกล้ 1 อย่างรวดเร็ว มากกว่าฟังก์ชันโลจีสติคและฟังก์ชันโพรบิทที่ปรับแล้ว



รูปที่ 2.5 กราฟฟังก์ชันการแจกแจงสะสมของตัวแบบความน่าจะเป็นเชิงเส้น ตัวแบบโลจิสติค ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก

Fahrmeir , L. and G.Tutz. (1994).

3. ในด้านการคำนวณ ตัวแบบโพรบิทและตัวแบบโลจิสติคจะให้ผลใกล้เคียงกัน โดยตัวแบบโพรบิทอยู่บนรากฐานของการแจกแจงสะสมแบบปกติซึ่งมีแนวคิดและพื้นฐานทางทฤษฎีสนับสนุนอย่างมีเหตุมีผล แต่ตัวแบบโพรบิทมีการคำนวณที่ยุ่งยาก กล่าวคือในการกำหนดค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจจะต้องติดอยู่ในรูปแบบของการอินทิเกรตเสมอ จึงทำให้ไม่สะดวก และใช้เวลาในการคำนวณด้วยคอมพิวเตอร์นานกว่า แต่ตัวแบบโลจิสติคคำนวณได้ง่ายไม่ติดอยู่ในรูปเครื่องหมายการอินทิเกรต ทำให้สามารถประมาณค่าได้ แต่ถ้าตัวแปรตอบสนองเป็นตัวแปรตอบสนองที่มีหลายลักษณะแล้ว ผลการคำนวณของตัวแบบทั้งสองจะแตกต่างกันมาก (Judge และคณะ,1988)

ตารางที่ 2.1 : ค่าเฉลี่ยและความแปรปรวนของฟังก์ชันตอบสนอง

Response function F	Mean	Variance
linear	0.5	1/12
probit	0	1
logistic	0	$\pi^2/3$
Complementary log-log	-0.5772	$\pi^2/6$

2.6 ส่วนประกอบของ GLM

GLM เป็นตัวแบบที่ขยายจากตัวแบบเชิงเส้นแบบคลาสสิค (classical linear model) โดยส่วนประกอบแรกที่เกี่ยวกับตัวแปรเชิงสุ่มนั้น นอกจากมีการแจกแจงแบบปกติแล้วยังสามารถขยายไปสู่การแจกแจงในกลุ่มเอกซ์โปเนนเชียล (exponential family) ได้และส่วนประกอบของ **link function** นอกจากจะใช้ **identical link** แล้ว ก็ยังสามารถขยายให้ใช้กับฟังก์ชันตัวเชื่อมอื่นๆอีกหลายแบบที่เป็นฟังก์ชันแบบ **monotonic differentiable function** ใดๆก็ได้ ส่วนประกอบทั้ง3 ส่วนมีรายละเอียดของแต่ละส่วนประกอบดังนี้

ส่วนประกอบที่ 1 ของ **GLM** คือส่วนประกอบที่เกี่ยวข้องกับสมมติฐานของการแจกแจงของตัวแปรเชิงสุ่ม (Y) ที่เป็นตัวแปรตอบสนอง สมมติว่าค่าสังเกตจาก Y มีขนาด n หน่วยที่เป็นอิสระต่อกัน นั่นคือ $Y = (y_1, \ldots, y_n)$ แต่ละส่วนประกอบของ Y คือ y_i , $i = 1, \ldots, n$ มีการ**แจกแจงในกลุ่มเอกซโปเนนเชียล** ซึ่งอยู่ในรูปของ

$$f_y(y;\theta,\phi) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)] \qquad (2.17)$$

โดยที่ a(.), b(.) และ c(.)แทนฟังก์ชันต่างๆเมื่อทราบ ϕ แล้ว (2.17) คือ ตัวแบบหนึ่งในกลุ่มเอกซ์โปเนนเชียลที่มีพารามิเตอร์ θ แต่ถ้าไม่ทราบ ϕ (2.17) อาจเป็นหรือไม่เป็นตัวแบบหนึ่งในกลุ่มเอกซ์โปเนนเชียลที่มีพารามิเตอร์ 2 ตัว (θ,ϕ) สำหรับพารามิเตอร์ θ เรียกว่า **natural parameter** ส่วน ϕ มักเรียกว่า **dispersion parameter** และฟังก์ชัน $a(\phi)$ มักจะมีรูปแบบเป็น

$$a(\phi) = \phi / w_i$$

โดยที่ w_i แทนน้ำหนักที่ทราบค่า เช่น เมื่อ $\bar{y}_i$ แทนค่าเฉลี่ยของ n_i หน่วยที่เป็นอิสระต่อกัน จะใช้โดยทั่วไปว่า $w_i = n_i$ และเพื่อให้เกิดความเข้าใจในตัวแบบ (2.17) ได้ชัดเจนขึ้นจะยกตัวอย่าง **การแจกแจงแบบปกติ** ซึ่งสามารถเขียนให้อยู่ในรูปแบบของ (2.17) ได้ดังนี้

$$\begin{aligned} f_y(y;\theta,\phi) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\{-(y-\mu)^2/2\sigma^2\} \\ &= \exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}\left(y^2/\sigma^2 + \log(2\pi\sigma^2)\right)\right\} \end{aligned}$$

โดยที่ $\theta = \mu$, $b(\theta) = \theta^2/2$, $a(\phi) = \phi = \sigma^2$ และ $c(y,\phi) = -1/2\left(y^2/\sigma^2 + \log(2\pi\sigma^2)\right)$

ทำนองเดียวกันกับการแจกแจงแบบปกติ **การแจกแจงแบบปัวส์ซอง** (Poisson) การแจกแจง**แบบทวินาม** (binomial) และการแจกแจงใน **กลุ่มเอกซ์โปเนนเชียลอื่นๆ** ก็สามารถเขียนให้อยู่ในรูปของ (5.1) ได้เช่นกัน และในกรณีที่ ϕ เป็นค่าคงที่ที่ทราบค่า (2.17) จะอยู่ในรูปของ (2.18) คือ

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp\{y_i Q(\theta)\} \qquad (2.18)$$

โดยที่ $Q(\theta)$ ใน (2.18) คือ $\theta/a(\phi)$ ใน (2.17)

$a(\theta)$ ใน (2.18) คือ $\exp\{-b(\theta)\}/a(\phi)$ ใน (2.17)

$b(y)$ ใน (2.18) คือ $\exp\{c(y,\phi)\}$ ใน (2.17)

จะเห็นว่าตัวแบบ (2.17) มีรูปแบบทั่วไป ที่สามารถนำไปใช้ประโยชน์กับการแจกแจงหลายรูปแบบ โดยเฉพาะสำหรับกลุ่มพารามิเตอร์ 2 ตัว (two –parameter families) เช่นการแจกแจงแบบปกติและการแจกแจงแบบแกมม่า (gamma) ซึ่ง ϕ จะเป็นพารามิเตอร์ของความคลาดเคลื่อน (nuisance parameter) ส่วนการแจกแจงสำหรับกลุ่มที่มีพารามิเตอร์ตัวเดียว (one–parameter families) เช่น การแจกแจงแบบปัวส์ซอง แบบทวินาม ไม่จำเป็นต้องใช้เทอม ϕ

ส่วนประกอบที่ 2 ของ GLM คือ ส่วนประกอบแบบมีระบบ ทำหน้าที่เชื่อมเวกเตอร์ η โดยที่ $\eta = (\eta_1,...,\eta_N)'$ กับเซตของตัวแปรอธิบาย ให้มีรูปแบบเชิงเส้นดังนี้

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \quad ; \quad \eta_j = \sum_j \beta_i x_{ij} \quad i = 1,...,p \;\; j = 1,...,N$$

โดยที่ **X** แทนเมทริกซ์ของตัวแปรอธิบายที่ประกอบด้วยค่าสังเกตขนาด N อาจเรียก X ว่า design matrix ที่มีขนาด $(N \times p)$

$\boldsymbol{\beta}$ แทนเวกเตอร์ของพารามิเตอร์ $(\beta_1,...,\beta_p)'$

$\boldsymbol{\eta}$ แทนตัวพยากรณ์เชิงเส้น (linear predictor)

ส่วนประกอบที่ 3 ของ GLM คือ link function ต่างๆสำหรับเชื่อมส่วนประกอบเชิงสุ่มและส่วนประกอบแบบมีระบบเข้าด้วยกัน เช่น

ให้ $\mu_j = E(Y_j)$, $j = 1,..,N$

$\therefore$ μ_j จะเกี่ยวข้องกับ η_j ในรูปฟังก์ชันของ $\eta_j = g(\mu_j)$

โดยที่ g แทนฟังก์ชันแบบ **monotonic differentiable function** ดังนั้นตัวแบบที่จะต้องการเชื่อมระหว่างค่าเฉลี่ยของค่าสังเกตของ Y กับตัวแปรอธิบาย คือ

$$g(\mu_j) = \sum_i \beta_i x_{ij} \quad , \; i=1,\dots,p \quad j=1,\dots,N$$

โดยที่ p แทนจำนวนตัวแปรอธิบาย

ถ้า $g(\mu)=\mu$ จะได้ว่า $\eta_j = \mu_j$ คือ **identity link** หรือเรียกอีกอย่างหนึ่งว่า **canonical link** โดยมีการแปลงค่าเฉลี่ยให้อยู่ในเทอมของพารามิเตอร์นั่นคือ

$$g(\mu_j) = Q(\theta_j)$$

และ
$$Q(\theta_j) = \sum_i \beta_i x_{ij} \qquad i=1,\dots,p$$

สรุปว่า GLM เป็นตัวแบบเชิงเส้นสำหรับค่าเฉลี่ยที่แปลงแล้วของตัวแปรซึ่งมีการแจกแจงอยู่ในกลุ่มเอกซ์โปเนนเชียล (McCullage ,1983)

2.7 การประมาณพารามิเตอร์สำหรับตัวแบบ GLM

ส่วนประกอบที่ 1 เกี่ยวข้องกับฟังก์ชันความน่าจะเป็นของ Y ใน (2.17) ซึ่งมีรูปแบบเป็น

$$f_y(y;\theta,\phi)=\exp[\{y\theta - b(\theta)\}/a(\phi)+c(y,\phi)] \tag{2.19}$$

ถ้าให้ $l_i = l(\theta,\phi;y) = \log f(y_i;\theta_i,\phi)$ แทนฟังก์ชัน log-likelihood จากหน่วยที่ i ของ $f_y(y_i;\theta_i,\phi)$ การคำนวณ **ค่าเฉลี่ย** และ **ความแปรปรวน** ของตัวแปรตัวแปรเชิงสุ่ม สามารถทำได้ดังนี้

จาก
$$l_i = \log f(y_i;\theta_i,\phi)$$

$$\therefore \quad l_i = [\{y_i\theta_i - b(\theta_i)\}/a(\phi)+c(y_i,\phi)]$$

$$\frac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi) \tag{2.20}$$

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = -b''(\theta_i)/a(\phi) \tag{2.21}$$

โดยที่ $b'(\theta_i)$ และ $b''(\theta_i)$ แทน derivatives ของ $b(\theta)$ ครั้งที่ 1และที่ 2 ตามลำดับ ผลลัพธ์จาก (2.20) และ(2.21) คือ (2.22) และ (2.23) ตามลำดับ

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{2.22}$$

และ

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0 \qquad (2.23)$$

จาก (2.20) และ (2.21) จะได้ค่าเฉลี่ย $E(Y_i)$ จาก

$$\{\mu - b'(\theta_i)\}/a(\phi) = 0$$

$$\therefore \quad E(Y_i) = \mu = b'(\theta_i)$$

ทำนองเดียวกันจาก (2.22) และ (2.23) จะได้ความแปรปรวนจาก

$$-\frac{b''(\theta_i)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)} = 0$$

$$\left[b''(\theta_i)/a(\phi) = E\{Y - b'(\theta_i)/a(\phi)\}^2 = Var(Y_i)/\{a(\phi)\}^2\right]$$

$$\therefore \quad Var(Y_i) = b''(\theta_i)a(\phi)$$

หมายเหตุ : จะเห็นว่าความแปรปรวนของ Y เป็นผลคูณของฟังก์ชัน $b''(\theta_i)$ และฟังก์ชัน $a(\phi)$ โดยฟังก์ชัน $b''(\theta_i)$ ขึ้นอยู่กับ**แคนอนิคอลพารามิเตอร์** (canonical parameter) θ เท่านั้น (คือ เป็นเทอมที่ขึ้นอยู่กับค่าเฉลี่ย) และเรียกว่า **ฟังก์ชันความแปรปรวน** การที่ขึ้นอยู่กับค่าเฉลี่ยด้วย จึงอาจเขียนเป็น $V(\mu)$ ส่วนฟังก์ชัน $a(\phi)$ เป็นอิสระจาก θ และโดยทั่วไปมักเขียนอยู่ในรูปของ

$$a(\phi) = \phi/w$$

โดยที่ ϕ แทน σ^2 จึงเรียก ϕ ว่า **dispersion parameter** ซึ่งต้องมีค่าคงที่ ณ ค่าสังเกตต่างๆ ส่วน w แทนน้ำหนักที่ทราบค่ามาก่อน ซึ่งอาจมีค่าต่างกันระหว่างค่าสังเกตหนึ่งๆ ของ Y ดังนั้น สำหรับตัวแบบที่ค่าสังเกต y_i แต่ละหน่วยเป็นค่าเฉลี่ยจาก m_i จะได้ว่า

$$a(\phi) = \sigma^2/m$$

$$\text{และ} \quad w = m \quad \text{หรือ} \quad w_i = m_i$$

ส่วนประกอบที่ 2 ซึ่งแสดงความสัมพันธ์ระหว่างพารามิเตอร์ η กับตัวแปรอธิบายต่างๆ โดยใช้ตัวแบบเชิงเส้นของ

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

หรือ

$$\eta_i = \sum_j \beta_j x_{ij} \quad i = 1,\ldots,N \;\; j = 1,\ldots,p$$

โดยที่ $\mathbf{X}$ แทนเมทริกซ์ขนาด $(N \times p)$

$\boldsymbol{\beta}$ แทนเวกเตอร์ของพารามิเตอร์ $(p \times 1)$

$\boldsymbol{\eta}$ แทนเวกเตอร์ $(N \times p)$ ตัวพยากรณ์เชิงเส้น (linear predictor)

ส่วนประกอบที่ 3 คือ ลิ้งค์ฟังก์ชัน (link function) สำหรับเชื่อมโยงค่าคาดหวังของ Y_i คือ μ_i กับตัวพยากรณ์เชิงเส้น โดยใช้

$$\eta_i = g(\mu_i)$$

โดยที่ g แทนฟังก์ชันแบบ **monotonic differentiable function**

ดังนั้น ตัวแบบ GLM จะเกี่ยวข้องกับค่าคาดหวังของตัวแปรเชิงสุ่ม Y หรือตัวแปรตอบสนองกับตัวแปรอธิบายในรูปแบบดังนี้

$$g(\mu_i) = \sum_j \beta_j x_{ij}$$

สำหรับฟังก์ชัน g โดยเฉพาะ $g(\mu_i) = \theta_i$ เรียกว่า **แคนนอนิคอลลิงค์** (canonical link) ดังนั้นจะได้ความสัมพันธ์ระหว่างพารามิเตอร์ θ_i และตัวพยากรณ์เชิงเส้นที่อยู่ในรูปแบบของ

$$\theta_i = \sum_j \beta_j x_{ij}$$

และเนื่องจาก $\mu_i = b'(\theta_i)$ ผลลัพธ์ คือ แคนนอนิคอลลิงค์ จะเป็นส่วนกลับ(inverse) ของฟังก์ชัน $b'(\theta_i)$ นั่นเอง

โดยสรุปการประมาณค่าพารามิเตอร์ของตัวแบบ GLM จึงมุ่งประมาณค่าของ β_j , $j=1,\ldots,p$ ตลอดจนค่าเฉลี่ยและค่าความแปรปรวนของ Y ดังที่กล่าวแล้ว สำหรับ **การประมาณค่าพารามิเตอร์** β_j , $j=1,\ldots,p$ อาจอาศัย **สมการปกติของฟังก์ชันภาวะน่าจะเป็น** ดังนี้

จาก log-likelihood สำหรับค่าสังเกต 1 ค่า มีรูปแบบเป็น

$$l_i = [\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i,\phi)]$$

สิ่งที่ต้องการคือนิพจน์สำหรับ $\dfrac{\partial l_i}{\partial \beta_j}$ จึงควรอาศัย **กฎลูกโซ่** (chain rule) ในรูปแบบดังต่อไปนี้

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{2.24}$$

แต่จาก $\frac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi)$ และ $\mu_i = b'(\theta_i)$, $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = Var(Y_i)/a(\phi)$ หรือ $a(\phi)b''(\theta_i) = Var(Y_i)$

$$\therefore \quad \frac{\partial l_i}{\partial \theta_i} = \frac{(y_i - \mu)}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{Var(Y_i)}$$

และจาก $\eta_i = \sum_j \beta_j x_{ij}$ ดังนั้น $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

แทนค่าใน(2.24) จะได้ว่า

$$\therefore \quad \frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \cdot \frac{a(\phi)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} . x_{ij}$$

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} . x_{ij} \tag{2.25}$$

เนื่องจาก log-likelihood สำหรับค่าสังเกต N ค่า จะอยู่ในรูปของผลบวกของ $\log f(y_i;\theta_i,\phi)$ นั่นคือ

$$L = \sum_i l_i = \sum_i \log f(y_i;\theta_i,\phi)$$

ดังนั้น **สมการภาวะน่าจะเป็นปกติ** (likelihood equation) หรือ **สมการปกติ** (normal equation) จาก (2.25) คือ

$$\sum_i^N \frac{(y_i - \mu_i)X_{ij}}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad , \quad j = 1,\dots,p \tag{2.26}$$

แต่เนื่องจากสมการภาวะน่าจะเป็นข้างต้นเป็น **ฟังก์ชันไม่เป็นเชิงเส้น** (nonlinear functions) ของ β การแก้สมการเพื่อหาตัวประมาณของ β จึงต้องใช้วิธีย้อนซ้ำ (iterative methods) ซึ่งสำหรับ GLM ใช้วิธีที่เรียกว่า **Fisher scoring** เป็นวิธีการคล้ายกับวิธีของ **Newton Raphson**

2.8 การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ (iterative maximum-likelihood estimation)

ความหมายของวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ

การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ (Agresti 1990 : 445-453) เป็นวิธีที่ใช้ในการประมาณค่าของพารามิเตอร์ของตัวแบบในกลุ่มเอ็กโพเนนเชียล (exponential family models) โดยสืบเนื่องมาจาก Nelder, J.A. และ Wedderburn, R.W.M.(1972) ได้ขยายวิธีการประมาณค่าแบบย้อนซ้ำ (iterative estimation) วิธีหนึ่งที่เรียกว่า วิธีฟิชเชอร์-สกอริง (Fisher scoring) ให้ใช้ควบคู่กับวิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood) และใช้สำหรับการประมาณค่าพารามิเตอร์ของตัวแบบเชิงเส้นที่วางนัยทั่วไป (generalized linear models) ซึ่งเป็นตัวแบบที่หมายรวมถึงตัวแบบเชิงเส้นที่มีพื้นฐานของการแจกแจงแบบปกติและตัวแบบอื่นๆ ที่มีการแจกแจงในกลุ่มของเอ็กโพเนนเชียลด้วย

การประมาณค่าพารามิเตอร์ของตัวแบบเชิงเส้นทั่วไปในอดีต นิยมใช้วิธีภาวะน่าจะเป็นสูงสุดและวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weighted least square) อย่างไรก็ตามในหลายสถานการณ์พบว่าไม่สามารถใช้วิธีการประมาณค่าดังกล่าวได้โดยตรง เนื่องจากสมการปกติที่พบอาจไม่มีลักษณะเชิงเส้น หรือมีรูปแบบไม่เชิงเส้น (nonlinear) ในเทอมของพารามิเตอร์ การแก้สมการเหล่านี้จึงจำเป็นต้องมีการคำนวณเพิ่มเติมด้วยวิธีการย้อนซ้ำเชิงตัวเลข (numerical iteration) ซึ่งวิธีการย้อนซ้ำ (iterative procedures) หลายวิธี และเมื่อนำมาใช้ร่วมกับวิธีภาวะน่าจะเป็นสูงสุดด้วย ทำให้เป็นวิธีที่มีประสิทธิภาพมากขึ้น และเป็นที่นิยมใช้ในปัจจุบัน โดยสามารถใช้โปรแกรมสำเร็จรูปต่างๆมาช่วยคำนวณได้ เช่น GLIM SPSS/FW SAS ฯลฯ

การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำที่จะกล่าวต่อไปนี้มีกระบวนการย้อนซ้ำ (iterative process) ที่เริ่มต้นจากวิธีภาวะน่าจะเป็นสูงสุดก่อน ส่วนกระบวนการปกติภายใต้การย้อนซ้ำจะอยู่ในลักษณะสมการปกติของวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก ที่ทำให้ค่าของการถ่วงน้ำหนัก (weight matrix) เปลี่ยนไปในแต่ละรอบ (cycle) ของการย้อนซ้ำต่างๆ กระบวนการย้อนซ้ำนี้จะจบลงเมื่อค่าประมาณของพารามิเตอร์ลู่เข้า (converge) สู่ค่าประมาณภาวะน่าจะเป็นสูงสุด (maximum likelihood estimates) หรือทำให้ค่าผลต่างของค่าประมาณจากกระบวนการย้อนซ้ำนั้น มีค่าน้อยหรือเล็กเพียงพอ (sufficient small) จึงเรียกกระบวนการดังกล่าวว่า การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ

การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ ใช้ได้สำหรับการประมาณค่าพารามิเตอร์ทั้งของตัวแบบเชิงเส้น ตัวแบบเชิงเส้นที่วางนัยทั่วไป และตัวแบบไม่เชิงเส้นอื่นๆ โดยสามารถใช้หลักเกณฑ์ของวิธีภาวะน่าจะเป็นสูงสุด ร่วมกับการใช้กระบวนการย้อนซ้ำจากวิธีหนึ่งวิธีใดใน 3 วิธีต่อไปนี้

1.วิธีนิวตัน – รัฟสัน (Newton – Raphson mathod)

2.วิธีฟิชเชอร์ – สกอริง (Fisher – scoring method หรือ Fisher's scoring หรือ the method of scoring)

3.วิธีเดมมิ่ง – สตีเฟน IPE (Deming – Stephen Iterative Proportional Fitting method หรือเรียกย่อๆว่า IPE)

ในกรณีที่ตัวแปรต่างๆของตัวแบบมีการแจกแจงแบบปกติ และสมการมีลักษณะของตัวแบบเชิงเส้น การประมาณค่าพารามิเตอร์ของตัวแบบจะไม่ต้องมีการย้อนซ้ำ กล่าวคือ กระบวนการย้อนซ้ำจะมีเพียงหนึ่งรอบเท่านั้น ค่าประมาณที่ได้จึงเป็นค่าเดียวกันกับค่าที่ได้จากวิธีภาวะน่าจะเป็นสูงสุดโดยตรง หรืออาจใช้วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักในอดีต

วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood method or ML)

ให้ $Y_1, Y_2, ..., Y_N$ แทนตัวแปรสุ่ม N ที่มี**ฟังก์ชันความหนาแน่นของความน่าจะเป็นร่วม (joint probability density function)** คือ

$$f(Y;\theta) \;=\; f(Y_1,...,Y_N;\theta_1,...,\theta_p)$$

ขึ้นอยู่กับพารามิเตอร์ $\theta_1, ..., \theta_p$ โดยที่ Y แทน $(Y_1,...,Y_N)'$ และ θ แทน $(\theta_1,...,\theta_p)'$ ส่วน p แทนจำนวนพารามิเตอร์ สำหรับฟังก์ชัน $f(Y;\theta)$ นั้น Y เป็นตัวแปรเชิงสุ่ม และ θ เป็นค่าคงที่ ในกรณีที่มีการสังเกตค่าของตัวแปรเชิงสุ่ม Y จำนวน N ค่าที่เป็นอิสระต่อกัน คือ $y_1,...,y_N$ ภายใต้พารามิเตอร์ θ **ฟังก์ชันภาวะน่าจะเป็น** (likelihood function) คือ

$$\begin{aligned} f(Y;\theta) &= f(y_1,...,y_N;\theta_1,...,\theta_p) \\ &= f(y_1;\theta_1)f(y_2;\theta_2)...f(y_p;\theta_p) \end{aligned}$$

เป็นที่น่าสังเกตว่า y ในฟังก์ชันภาวะน่าจะเป็นจะเหมือนกับ Y ในฟังก์ชันความหนาแน่นของความน่าจะเป็นร่วม เพียงแต่ y เป็นค่าสังเกตของ Y

ให้ $\quad L = \log\ f(y;\theta) = \sum_i \log\ f(y_i;\theta) = \sum_i l_i$

ให้ Ω แทนเซตที่เป็นไปได้ทั้งหมดของเวกเตอร์ของพารามิเตอร์ θ

ดังนั้นค่าประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimate หรือ MLE) ของ θ คือ $\hat{\theta}$ ที่ทำให้ฟังก์ชันภาวะน่าจะเป็น $f(y;\theta)$ มีค่าสูงสุด นั่นคือ

$$f(y;\hat{\theta}) \geq f(y;\theta) \qquad \text{สำหรับทุก } \theta \text{ ใน } \Omega$$

ค่าประมาณภาวะน่าจะเป็นสูงสุดหรือ $\hat{\theta}$ นี้ มีค่าที่ทำให้ $L(\theta)$ สูงสุดด้วย เนื่องจากฟังก์ชันลอการิทึมเป็นฟังก์ชันที่มีผลในทางเดียวกัน (monotonic) ดังนั้น

$$\log\ f(y;\hat{\theta}) \geq \log\ f(y;\theta) \qquad \text{สำหรับทุก } \theta \text{ ใน } \Omega$$

หรือ
$$L(\hat{\theta}) \geq L(\theta) \qquad \text{สำหรับทุก } \theta \text{ ใน } \Omega$$

โดยทั่วไป นิยมใช้ฟังก์ชันของ log-likelihood มากกว่าฟังก์ชันของ likelihood โดยตรง เนื่องจากช่วยให้การคำนวณสะดวกขึ้น สำหรับตัวประมาณแบบ MLE สามารถหาได้จากดิฟเฟอเรนชิเอท (differentiating) ฟังก์ชัน L เทียบกับพารามิเตอร์ทีละตัว แล้วแก้สมการปกติพร้อมกัน คือ

$$\frac{\partial}{\partial\theta_j}L = 0 \qquad \text{สำหรับ } j = 1,\ldots,p$$

การตรวจว่าฟังก์ชัน L มีค่าสูงสุดหรือไม่ สามารถทำได้โดยการหาเมทริกซ์ของ second derivatives ของ L คือ $\frac{\partial^2 L}{\partial\theta_j\partial\theta_k}$ ว่ามีค่าเป็น negative definite ณ ค่า $\hat{\theta}$ หรือไม่ ตัวอย่างของกรณีนี้เช่น ถ้ามีพารามิเตอร์ θ เพียงตัวเดียว การตรวจสอบฟังก์ชัน L ว่ามีค่าสูงสุดหรือไม่ จะสามารถทำได้โดยการหาค่าของ $\frac{\partial^2 L}{\partial\theta^2}$ ว่ามีค่าเป็นลบ ณ ค่าของ $\hat{\theta}$ หรือไม่

คุณสมบัติที่สำคัญอย่างหนึ่งของตัวประมาณภาวะน่าจะเป็นสูงสุด หรือ MLE คือคุณสมบัติที่เรียกว่า " invariance property of MLE " คือถ้า $f(\theta)$ เป็นฟังก์ชันใดๆ ของพารามิเตอร์ θ แล้วจะได้ว่า $f(\hat{\theta})$ เป็นตัวประมาณภาวะน่าจะเป็นสูงสุดของ $f(\theta)$ ด้วย โดยที่ $\hat{\theta}$ เป็น MLE ของ θ ผลลัพธ์ดังกล่าวนี้คือ θ คุณสมบัติของ invariance ข้างต้น และด้วยคุณสมบัติของ invariance นี้ ทำให้สามารถหาตัวประมาณภาวะน่าจะเป็นสูงสุดของฟังก์ชันของ MLE ได้

นอกจากนี้ตัวประมาณภาวะน่าจะเป็นสูงสุดยังมีคุณสมบัติของ consistency, sufficiency และ asymptotic efficiency ด้วย

2.9 วิธีนิวตัน-รัฟสัน และวิธีฟิชเชอร์ – สกอริง สำหรับตัวแบบเชิงเส้นที่วางนัยทั่วไป

จากการหาค่าประมาณพารามิเตอร์แบบย้อนซ้ำทั้งวิธี Newton-Raphson และวิธี Fisher scoring สามารถเริ่มจากวิธีภาวะน่าจะเป็นสูงสุดภายใต้ลักษณะการแจกแจงของตัวแปรของตัวแบบที่สนใจ และในทำนองเดียวกันเมื่อพิจารณาถึงตัวแบบเชิงเส้นที่วางนัยทั่วไปสำหรับค่าสังเกตที่เป็นอิสระต่อกันจำนวน N นั้น ฟังก์ชัน log-likelihood ของตัวแปรเชิงสุ่ม Y_i คือ

$$L = \sum_i \log f(y_i;\theta_i,\phi) = \sum_i l$$

โดยที่ θ แทน nature parameter ที่ขึ้นกับ model-parameter เช่น β

ส่วน $a(\phi)$ แทน dispersion parameter และ

$$f_y(y_i;\theta_i,\phi) = \exp[\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i,\phi)]$$

ดังนั้น

$$l_i = \log\exp[\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i,\phi)]$$

$$\therefore \quad l_i = [\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i,\phi)]$$

เนื่องจากตัวแบบเชิงเส้นที่วางนัยทั่วไปประกอบด้วยส่วนประกอบ 3 ส่วน ได้แก่

1. ส่วนประกอบเชิงสุ่ม คือ $f(y_i;\theta_i,\phi)$
2. ส่วนประกอบแบบมีระบบ คือ $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$
3. ส่วนประกอบ "link function" คือ $\eta_i = g(\mu_i)$ เช่น $\eta = \mu$ หรือ $E(Y)$ ที่อยู่ในกรณีของ Normal หรือ Poisson

$$\therefore \ g(\mu_i) = \mathbf{X}\boldsymbol{\beta} = \sum_j \beta_j x_{ij} = \eta_i$$

โดยที่ฟังก์ชัน g ซึ่งทำให้ $g(\mu_i) = \theta_i$ เรียกว่า "canonical link function" และ g เป็น monotonic differentiable function

$$\therefore \quad \theta_i = \sum_j \beta_j x_{ij}$$

นั่นคือพารามิเตอร์ θ_i ของส่วนประกอบเชิงสุ่มข้างต้นขึ้นอยู่กับพารามิเตอร์ β

ดังนั้นการหา MLE จากฟังก์ชัน log-likelihood หรือ L ข้างต้น สามารถใช้วิธี differentiation แบบกฎลูกโซ่ (chain rule) คือ

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{2.27}$$

โดยที่ $\frac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi)$

แต่จาก Cox & Hinkley (1974) พบว่า

1. $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ และ $-E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left(\frac{\partial l}{\partial \theta}\right)^2$

ดังนั้น

$$E(Y_i) = \mu = b'(\theta_i)$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{(y_i - \mu)}{a(\phi)} \tag{2.28}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

และ $$\frac{\partial^2 l_i}{\partial \theta^2_i} = \frac{-b''(\theta)}{a(\phi)}$$

2. $b''(\theta_i)/a(\phi) = E[(Y-\mu)/a(\phi)]^2 = Var(Y_i)/[a(\phi)]^2$

ดังนั้น

$$b''(\theta_i) = Var(Y_i)/a(\phi)$$

$$\therefore \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{Var(Y_i)} \tag{2.29}$$

เนื่องจาก $$\eta_i = \sum_j \beta_j x_{ij}$$

$$\therefore \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \tag{2.30}$$

แทน (2.28)-(2.30) ใน (2.27) จะได้

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \cdot \frac{a(\phi)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} . x_{ij}$$

$$\therefore \frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} . x_{ij}$$

ดังนั้น สมการภาวะน่าจะเป็นปกติ (likelihood equation) หรือ สมการปกติ (normal equation) คือ

$$\frac{\partial l_i}{\partial \beta_j} = U_j = \sum_i^N \frac{(y_i - \mu_i)X_{ij}}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad , \quad j = 1,\dots,p \qquad (2.31)$$

อย่างไรก็ตาม (2.31) เป็น nonlinear function ของ β ดังนั้นการแก้สมการ (2.31) จำเป็นต้องใช้การย้อนซ้ำของ Newton Raphson หรือ Fisher scoring มาช่วย โดยอัตราของ convergence ขึ้นอยู่กับ "information matrix" หรือ $E\left[-\partial^2 L(\beta)/\partial \beta_i \partial \beta_j\right]$

$$\because E\left(\frac{\partial^2 l_i}{\partial \beta_a \partial \beta_b}\right) = -E\left[\left(\frac{\partial l_i}{\partial \beta_a}\right)\left(\frac{\partial l_i}{\partial \beta_b}\right)\right]$$

$$= -E\left[\frac{(Y_i - \mu_i)x_{ia}}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \left(\frac{(Y_i - \mu_i)x_{ib}}{Var(Y_i)}\right) \cdot \frac{\partial \mu_i}{\partial \eta_i}\right]$$

$$= \frac{-x_{ia}x_{ib}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

$$\therefore E\left(\frac{\partial^2 \mathbf{L}}{\partial \beta_a \partial \beta_b}\right) = -\sum_i^N \frac{x_{ia}x_{ib}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \qquad (2.32)$$

ถ้า "information matrix" มีสมาชิกเป็น $E\left[-\partial^2 L(\beta)/\partial \beta_a \partial \beta_b\right]$ และเขียนให้อยู่ในรูปของเมทริกซ์ คือ $\mathbf{M} = \mathbf{X'WX}$ โดยที่ $\mathbf{W}$ แทน main diagonal matrix ซึ่งมีสมาชิกเป็น

$$w_i = (\partial \mu_i / \partial \eta_i)^2 \ / Var(Y_i)$$

ในการประมาณค่าของ $\boldsymbol{\beta}$ ด้วยวิธี Newton-Raphson ทำโดยใช้วิธีการย้อนซ้ำครั้งที่ s ใดๆ จากสมการ ต่อไปนี้คือ

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{H}^{(s)}\right)^{-1}\mathbf{U}^{(s)}$$

โดยที่ $\mathbf{H}$ แทน เมทริกซ์ ซึ่งมีสมาชิกเป็น $-\partial^2 \mathbf{L}(\boldsymbol{\beta})/\partial \beta_a \partial \beta_b$ และ

$\mathbf{U}$ แทน เวกเตอร์ ซึ่งมีสมาชิกเป็น $-\partial \mathbf{L}(\boldsymbol{\beta})/\partial \beta_j$

ทั้ง $\mathbf{H}^{(s)}$ และ $\mathbf{U}^{(s)}$ คือค่าของ $\mathbf{H}$ และ $\mathbf{U}$ ในครั้งที่ s ณ การคำนวณค่าประมาณของ $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s)}$

$$\therefore \quad \boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left[\frac{\partial^2 l}{\partial \beta_a \partial \beta_b}\right]^{-1}_{\beta = \beta^{(s)}} \mathbf{U}^{(s)}$$

อนึ่งสำหรับเมทริกซ์ $\mathbf{H}$ นั้น อาจเขียนให้อยู่ในรูปของเมทริกซ์ $\mathbf{M}$ โดยที่ $\mathbf{M} = E(H)$ นั่นคือ ใช้สูตรสำหรับ Fisher scoring และจะได้ว่า

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{M}^{(s)}\right)^{-1}\mathbf{U}^{(s)}$$

หรือ
$$\mathbf{M}^{(s)}\boldsymbol{\beta}^{(s+1)} = \left(\mathbf{M}^{(s)}\right)\boldsymbol{\beta}^{(s)} + \mathbf{U}^{(s)} \tag{2.33}$$

โดยที่ $\mathbf{M} = \mathbf{X'WX}$ และมีสมาชิกเป็น $\mathrm{E}\left[-\partial^2 \mathrm{L}(\beta)/\partial\beta_a\partial\beta_b\right]$ ดังกล่าวแล้ว

$$\therefore \quad X'W^{(s)}X\beta^{(s+1)} = X'W^{(s)}X\beta^{(s)} + U^{(s)}$$

จะเห็นว่าสมการ(2.33) เป็นวิธีประมาณค่าภาวะน่าจะเป็นสูงสุดที่มีการย้อนซ้ำแบบ Fisher scoring ซึ่งมีการถ่วงน้ำหนักเมทริกซ์ $\mathbf{M}$ โดยที่ $\mathbf{M} = \mathbf{X'WX}$ หรือเปรียบเสมือนการประมาณแบบ IWLS (Iterative Weighted Least Squares) ซึ่งจะเห็นได้ชัดเจนจากด้านขวามือของ(2.33) ซึ่งอาจเขียนอยู่ในรูป

$$\sum_j\left[\sum_i \frac{x_{ia}x_{ij}}{\mathrm{Var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\beta_j^{(s)}\right] + \sum_j \frac{\left(y_i - \mu_i^{(s)}\right)x_{ia}}{\mathrm{Var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$$

โดยที่ μ_i และ $\left(\partial\mu_i/\partial\eta\right)$ คำนวณค่า ณ $\boldsymbol{\beta}^{(s)}$

$$\therefore \quad \left(\mathbf{M}^{(s)}\right)\boldsymbol{\beta}^{(s)} + \mathbf{U}^{(s)} = \mathbf{X'W}^{(s)}\,\boldsymbol{\psi}^{(s)}$$

โดยที่ $\mathbf{W}^{(s)}$ คือ $\mathbf{W}$ ที่มีสมาชิกเป็น $(\partial\mu_i/\partial\eta_i)^2 \ /\mathrm{Var}(Y_i)$ ณ $\boldsymbol{\beta}^{(s)}$ และ $\boldsymbol{\psi}^{(s)}$ ที่มีสมาชิกเป็น

$$\begin{aligned}\psi_i^{(s)} &= \sum_j x_{ij}\beta_j^{(s)} + \left(y_i - \mu_i^{(s)}\right)\left(\frac{\partial\eta_i^{(s)}}{\partial\mu_i^{(s)}}\right)\\ &= \eta_i^{(s)} + \left(y_i - \mu_i^{(s)}\right)\left(\frac{\partial\eta_i^{(s)}}{\partial\mu_i^{(s)}}\right)\end{aligned} \tag{2.34}$$

ดังนั้นวิธี Fisher scoring ใน (2.33) เขียนใหม่ได้เป็น

$$\begin{aligned}\mathbf{M}^{(s)}\boldsymbol{\beta}^{(s+1)} &= \mathbf{X'W}^{(s)}\boldsymbol{\psi}^{(s)}\\ \left(\mathbf{X'W}^{(s)}\mathbf{X}\right)\boldsymbol{\beta}^{(s+1)} &= \mathbf{X'W}^{(s)}\boldsymbol{\psi}^{(s)}\end{aligned} \tag{2.35}$$

ซึ่งมีรูปแบบของสมการปกติคล้ายวิธีกำลังสองน้อยที่สุดสำหรับตัวแบบเชิงเส้นที่มีตัวแปรตามเป็น $\boldsymbol{\psi}^{(s)}$ ผลลัพธ์ที่ได้ คือ

$$\boldsymbol{\beta}^{(s+1)} = \left(\mathbf{X}'\mathbf{W}^{(s)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(s)}\boldsymbol{\psi}^{(s)}$$

เป็นที่น่าสังเกตว่า เวกเตอร์ $\boldsymbol{\psi}^{(s)}$ ใน (2.34) และ (2.35) มีรูปแบบเชิงเส้นของ link function ณ $\boldsymbol{\mu}$ ที่คำนวณภายใต้ตัวแปร Y ณ $\mathbf{y}$ นั่นคือ

$$g(y) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu) = \eta_i + (y_i - \mu_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right) = \psi_i$$

จะเห็นได้ชัดเจนว่า กระบวนการย้อนซ้ำข้างต้นอยู่กับตัวแปรตาม $\boldsymbol{\psi}$ ซึ่งมีสมาชิกที่ i ประมาณด้วย $\boldsymbol{\psi}^{(s)}$ ของการย้อนซ้ำครั้งที่ s เป็นการถดถอยของ $\boldsymbol{\psi}^{(s)}$ บน $\mathbf{X}$ พร้อมถ่วงน้ำหนักด้วย $\mathbf{W}^{(s)}$ ซึ่งทำให้ได้ผลลัพธ์ตัวใหม่คือ ตัวประมาณของ $\boldsymbol{\beta}^{(s+1)}$ โดยตัวประมาณค่านี้จะให้ผลลัพธ์ใหม่ $\boldsymbol{\eta}^{(s+1)} = \mathbf{X}\boldsymbol{\beta}^{(s+1)}$ ตลอดจนตัวแปรตามใหม่คือ $\boldsymbol{\psi}^{(s+1)}$ ซึ่งใช้สำหรับรอบต่อไป

ดังนั้นตัวประมาณภาวะน่าจะเป็นสูงสุด คือ ลิมิตของ $\boldsymbol{\beta}^{(s)}$ ในขณะที่ $s \to \infty$ กล่าวอีกนัยหนึ่ง คือ ตัวประมาณภาวะน่าจะเป็นของตัวแบบเชิงเส้นที่วางนัยทั่วไป เป็นผลของการย้อนซ้ำที่ใช้การถ่วงน้ำหนักของกำลังสองน้อยที่สุด โดยเมทริกซ์ของการถ่วงน้ำหนักจะเปลี่ยนค่าไปทุกๆครั้งของการย้อนซ้ำ เรียกกระบวนการนี้ว่า **“IRLS process”** (คือ Iterative Reweighted Least Square process) กระบวนการย้อนซ้ำทั้งหมดจะเสร็จสิ้นต่อเมื่อค่าประมาณของ β ที่มีการย้อนซ้ำสองครั้งติดต่อกัน ให้ค่าผลต่างที่เล็กเพียงพอ หรือเป็นศูนย์

อนึ่งความแปรปรวนร่วมเมื่อใกล้อนันต์ของ $\hat{\boldsymbol{\beta}}$ อยู่ในรูปเมทริกซ์ ที่เป็นส่วนกลับของเมทริกซ์ $\mathbf{M}$ เราสามารถหาค่าประมาณของ $Cov(\boldsymbol{\beta})$ จาก $C\hat{o}v(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$ โดยที่ $\hat{\mathbf{W}}$ คำนวณจากค่า $\mathbf{W}$ ณ $\hat{\boldsymbol{\beta}}$ เช่น $w_i = (\partial\mu_i/\partial\eta_i)^2/\mathrm{Var}(Y_i)$

โดยสรุปตัวแบบเชิงเส้นที่วางนัยทั่วไป ในการประมาณค่าพารามิเตอร์ของตัวแบบไม่ว่าจะใช้วิธี Fisher scoring หรือใช้วิธี Newton –Raphson โดยตรง ทั้งสองวิธีนี้พบว่ามีสิ่งที่ร่วมกัน คือ กระบวนการย้อนซ้ำต่างก็อาศัยวิธีภาวะน่าจะเป็นสูงสุดก่อน แล้วใช้การประมาณค่าจากสมการปกติในลักษณะของวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก โดยมีกระบวนการสร้างตัวถ่วงน้ำหนักขึ้นใหม่ในทุกๆรอบของการย้อนซ้ำ จนกระทั่งได้ตัวประมาณที่ลู่เข้าและมีคุณสมบัติตามต้องการ เช่น ความพอเพียง ความคงเส้นคงวา ฯลฯ วิธีดังกล่าวนี้อาจเรียกรวมกันได้ว่า การประมาณค่าภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ หรือวิธี IRLS ก็ได้ โดยเป็นวิธีประมาณค่าที่นิยมใช้สำหรับตัวแบบเชิงเส้นที่วางนัยทั่วไป ซึ่งหมายรวมถึงตัวแบบการถดถอยโลจิสติค ตัวแบบโลจิท ตัวแบบล็อกลิเนียร์ ตัวแบบอื่นๆในกลุ่มเอ็กโปเนนเชียล และตัวแบบที่ไม่เชิงเส้นทั่วไปด้วย

2.10 วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method หรือ WLS)

ให้ $Y_1, Y_2, ..., Y_N$ แทนตัวแปรเชิงสุ่มที่มีค่าคาดหวังเป็น

$$E(Y_i) = \mu_i \quad \text{สำหรับ} \quad i = 1,2,...,N$$

ให้ μ_i's เป็นฟังก์ชันของพารามิเตอร์ $\beta_1,...,\beta_p$ $(p \le N)$ และต้องการประมาณค่าของ $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก โดยใช้ค่าประมาณแทนด้วย $\hat{\boldsymbol{\beta}} = (\beta_1,...,\beta_p)'$ ตามลำดับ

ให้ $Y_i = \mu_i + e_i$ สำหรับ $i = 1,2,...,N$

โดยที่ e_i แทนค่าความคลาดเคลื่อนเชิงสุ่ม

จากวิธีกำลังน้อยที่สุดแบบธรรมดา (Ordinary Least Squares method หรือ OLS) ซึ่งมีวิธีการคำนวณสำหรับการหาค่าประมาณของ $\boldsymbol{\beta}$ ที่ทำให้ผลรวมกำลังสองของ e_i ต่ำสุด คือ

ให้

$$\begin{aligned} SS &= \sum e_i^2 \\ &= \sum (Y_i - \mu_i(\boldsymbol{\beta}))^2 \\ &= (\mathbf{Y} - \boldsymbol{\mu})'(\mathbf{Y} - \boldsymbol{\mu}) \end{aligned}$$

โดยที่ $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)'$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)'$

โดยทั่วไปตัวประมาณของ $\boldsymbol{\beta}$ หรือ $\hat{\boldsymbol{\beta}}$ สามารถหาได้จากดิฟเฟอเรนซิเอท เทอม SS เทียบกับ β_j ของ $\boldsymbol{\beta}$ แล้วแก้สมการ

$$\frac{\partial SS}{\partial \beta_j} = 0 \qquad \text{สำหรับ} \quad j = 1,2,...,p$$

การตรวจสอบค่าต่ำสุดของ SS ที่ใช้ประมาณค่า $\boldsymbol{\beta}$ ทำได้โดยการตรวจสอบเมทริกซ์ของ second derivatives ว่าเป็น positive definite หรือไม่ ถ้าใช้การประมาณที่ได้ก็ตรงกับจุดประสงค์

ในทางปฏิบัติอาจมี Y_i บางตัว ที่มีค่าสังเกตที่เชื่อถือได้น้อย กล่าวคือ อาจมีความแปรปรวนมากกว่า Y_i's ตัวอื่นๆ กรณีเช่นนี้อาจจำเป็นต้องมีการถ่วงน้ำหนักในเทอมของ SS และใช้เทอม SS_W แทนเทอม SS โดยที่

$$SS_W = \sum_i V_i [Y_i - \mu_i(\beta)]^2$$

โดยที่ $$V_i = \frac{1}{[Var(Y_i)]}$$

นอกจากนี้ Y_i's ยังอาจไม่เป็นอิสระต่อกัน (คือ มีความสัมพันธ์กัน) กรณีเช่นนี้จึงควรใช้ $V = 1/\mathbf{W}$ เมื่อ $\mathbf{W}$ แทน Variance-Covariance matrix ของ Y_i's ดังนั้นวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก หรือ วิธี WLS สามารถคำนวณได้จากการทำให้ SS_W มีค่าต่ำสุด โดยที่

$$SS_W = (\mathbf{Y}-\boldsymbol{\mu})'\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$$

แต่เนื่องจาก $\boldsymbol{\mu}$ เป็นฟังก์ชันของ $\boldsymbol{\beta}$ ซึ่งในกรณีที่ μ_j เป็น linear combination ของ β_j เมื่อ $j = 1,2,\ldots,p$ เช่น $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ โดยที่ $\mathbf{X}$ แทนเมทริกซ์ $\mathbf{X}$ ที่มีมิติขนาด $(N \times p)$ แล้ว จะได้ว่า

$$SS_W = (\mathbf{Y}-\boldsymbol{\mu})'\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$$

และ $$\frac{\partial SS_W}{\partial \beta} = -2\mathbf{X}'\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$$

ดังนั้นค่าประมาณ $\hat{\boldsymbol{\beta}}$ จากวิธี WLS สามารถคำนวณจากสมการปกติ

$$\begin{aligned} \mathbf{X}'\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\beta}\mathbf{X}) &= \mathbf{0} \\ \mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{W}^{-1}\mathbf{Y} \end{aligned}$$

นั่นคือ $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}$

หมายเหตุ *:*

1. สามารถตรวจสอบเมทริกซ์ second derivatives ของ SS_W ว่าเป็น positive definite หรือไม่
2. วิธี WLS สามารถใช้ได้โดยไม่ต้องมีข้อตกลง (assumptions) เกี่ยวกับการแจกแจงของ Y_i นอกเหนือจากการกำหนดค่าคาดหวังและโครงสร้างของ Variance-Covariance ของ Y_i's เท่านั้น อย่างไรก็ตามถ้าต้องการอนุมารเกี่ยวกับ $\boldsymbol{\beta}$ จำเป็นต้องมีข้อตกลงเกี่ยวกับ Y_i's เพิ่มเติม
3. ส่วนวิธี ML จำเป็นต้องทราบการแจกแจงของ Y_i's เนื่องจากต้องกำหนดฟังก์ชัน joint probability density ของ Y_i's

ประโยชน์ของวิธี WLS

วิธีกำลังสองน้อยสุดแบบธรรมดา (OLS) วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) เป็นวิธีที่ใช้สำหรับกรณีที่ความแปรปรวนของ Y_i คงที่และไม่คงที่ หรือ Y_i ไม่เป็นอิสระต่อกัน ตามลำดับ ถ้าความแปรปรวน Y_i คงที่หรือเท่ากันทุก Y_i และเป็นอิสระต่อกันแล้ว วิธี OLS และ WLS จะให้ผลลัพธ์เท่ากัน ทั้งสองวิธีนี้ต่างเป็นทางเลือกของวิธีภาวะน่าจะเป็นสูงสุด (ML) ประโยชน์ของวิธี WLS อาจสรุปได้ 4 ประการ คือ

1. การคำนวณของวิธี WLS มีรูปแบบมาตรฐาน (standard form) ที่ไม่ซับซ้อนและสามารถนำไปประยุกต์สำหรับการแก้สมการในตัวแบบอื่นๆได้

2. การคำนวณของวิธี ML ที่นำไปประยุกต์กับกระบวนการย้อนซ้ำ (iterative process) นั้น อาศัยประโยชน์ของการย้อนซ้ำด้วยวิธี WLS โดยแต่ละรอบของการย้อนซ้ำจะมีการถ่วงน้ำหนักด้วยเทอมที่มีการเปลี่ยนค่าไปเรื่อยๆ เช่นการใช้วิธี Newton-Raphson และวิธี Fisher scoring ร่วมกับวิธี ML เพื่อประมาณพารามิเตอร์ในตัวแบบล็อกลิเนียร์ และตัวแบบเชิงเส้นที่วางนัยทั่วไป

3. สำหรับตัวแบบที่เหมาะสม ตัวประมาณของ WLS และ ML เป็นตัวประมาณที่สมมูลกัน และต่างก็เป็นตัวประมาณแบบ BAN (Best Asymptotical Normal) เมื่อตัวอย่างมีขนาดใหญ่ ตัวประมาณเหล่านั้นมีการแจกแจงเข้าใกล้การแจกแจงแบบปกติ และอัตราส่วนของความแปรปรวนของทั้งสองวิธีจะลู่เข้าสู่ค่า 1 ด้วย

4. สำหรับการประมาณค่าแบบจุด (point estimation) ของพารามิเตอร์ด้วยวิธี WLS ไม่จำเป็นต้องมีข้อตกลงเกี่ยวกับลักษณะการแจกแจงของ Y แต่การประมาณค่าแบบช่วง (interval estimation) และการทดสอบสมมติฐานเกี่ยวกับตัวแบบ จำเป็นต้องทราบการแจกแจงเพิ่มเติม

ข้อจำกัดของวิธี WLS เมื่อเทียบกับวิธี ML

การใช้วิธี WLS สำหรับข้อมูลเชิงกลุ่ม ต้องมีการประมาณความแปรปรวนร่วมมัลติโนเมียล (multinomial covariance structure) ของข้อมูลตัวอย่างในแต่ละชุดของตัวแปรอธิบาย (each setting of the explanatory variables) อนึ่งในกรณีนี้ ถ้าตัวแปรอธิบายเป็นแบบต่อเนื่อง วิธี WLS อาจไม่เหมาะสม เนื่องจากอาจมีเพียง 1 ค่าสังเกตในแต่ละ setting นั้น ในขณะที่วิธี ML สามารถทำได้ นอกจากนี้ในกรณีที่จำนวนกลุ่มของตัวแปรอธิบายมีมาก วิธี WLS ยิ่งไม่เหมาะสม เพราะมีปัญหาเกี่ยวกับจำนวนนับในเซลล์มีน้อยหรือเป็นศูนย์ ซึ่งปัญหาเหล่านี้ไม่มีผลกระทบต่อการใช้วิธี ML โดยถ้าเซลล์เป็นศูนย์ วิธี ML อาจแทนค่าคงที่ๆ มีค่าน้อยมากในเซลล์ เพื่อให้สามารถคำนวณค่าประมาณจากวิธี ML พร้อมกับใช้วิธีการย้อนซ้ำปรับค่าถ่วงน้ำหนักต่างๆ ต่างกับส่วนของวิธี WLS โดยตรงที่ ไม่มีการย้อนซ้ำ เมื่อแทนค่าในเซลล์ศูนย์ ด้วยค่าเช่น 0.5 หรือค่าน้อยกว่านี้อีกมากๆ อาจทำให้เทอมความแปรปรวนมีค่าสูงหรือต่ำผิดปกติ จนกระทั่งมีผลกระทบอย่างมาก (strong influence) ต่อการวิเคราะห์การถ่วงน้ำหนัก ทำให้อาจลดความเชื่อถือทั้งผลลัพธ์และข้อมูล กรณีเช่นนี้จึงควรใช้วิธี ML แทน นอกจากนี้โปรแกรมสำเร็จรูปทางสถิติต่างๆ เช่น SPSS/FW, GLIM, MINITAB ซึ่งมีวิธี ML อยู่ จึงมีประโยชน์ต่อการใช้ในหลายสถานการณ์ (วีรานันท์ ; 2544)

2.11 การประมาณแบบไคกำลังสองต่ำสุด(Minimum chi-square : MCS)

เมื่อข้อมูลเป็นกลุ่ม เราสามารถหาการประมาณใกล้เคียงกับการภาวะน่าจะเป็นสูงสุด (maximum likelihood : ML) โดยใช้เส้นถดถอยกำลังสองน้อยสุดแบบถ่วงน้ำหนักอย่างง่ายในเอ็มพิริเคิลโลจิท หรือ โพรบิท (empirical logit or probit)

ตามที่แสดงให้เห็นมาก่อนหน้านี้ กับกลุ่มหรือข้อมูลที่มีการวัดซ้ำ จำนวนของ "ความสำเร็จ" (y_i)และจำนวนขนาดตัวอย่าง (n_i) สามารถหาความน่าจะเป็นแบบเอ็มพิริเคิล (empirical probabilities) $\widetilde{p}_i = y_i/n_i$ การแจกแจงในปัจจุบันเป็นทางเลือกหนึ่งที่นำไปสู่วิธีการประมาณแบบภาวะน่าจะเป็นสูงสุด(maximum likelihood estimation : MLE) ที่ใช้สัดส่วนของตัวอย่าง (ความน่าจะเป็นแบบเอ็มพิริเคิล) นำไปสู่รูปแบบเอ็มพิริเคิลโลจิทและโพรบิท (empirical logit or probit) นี้เป็นวิธีการการประมาณแบบคงเส้นคงวา นี้เป็นปรากฏการอย่างหนึ่งเพราะคล้ายกับการถดถอยมาตรฐาน

เราเริ่มต้นจากตัวแบบเชิงเส้นที่มีความแปรปรวนไม่คงที่ (Heteroscedastic) เหมือนกับตัวแบบความน่าจะเป็นมาก อย่างไรก็ตาม ตัวแปรตามเป็นเอ็มพิริเคิลโลจิท หรือ โพรบิท (empirical logit or probit) หาได้จากการแปลงของความน่าจะเป็นแบบเอ็มพิริเคิล (empirical probabilities) การตอบสนองในทางตรงกันข้าม (ผกผัน) ของความแปรปรวนของความคลาดเคลื่อน(error variance) เป็นการใช้โครงสร้างของการถ่วงน้ำหนักในการถดถอยแบบ FGLS

วิธีไคกำลังสองต่ำสุด (minimum chi-square : MCS) เริ่มต้นที่ตัวแบบเชิงเส้นสำหรับการแปลงที่ยึดตามทฤษฎี การตอบสนองความน่าจะเป็นของประชากร p_i

$$g(p_i) = x_i'\boldsymbol{\beta} = \eta_i$$

เมื่อ $g(p_i)$แสดงถึงทฤษฎีของโลจิท หรือโพรบิท ก่อนหน้านี้เป็นที่ทราบกันว่า $g(\cdot)$เป็นลิงค์ (link) ฟังก์ชันที่มาจากตัวแบบเชิงเส้นใน $\boldsymbol{\beta}$

เราสามารถแสดงออกได้ว่านี้เป็นตัวแบบถดถอยเชิงเส้นที่มีความแปรปรวนต่างกัน

$$g(\widetilde{p}_i) = x_i'\boldsymbol{\beta} + \varepsilon_i$$

โดยที่ $\widetilde{p}_i = y_i/n_i$ และ $\varepsilon_i = \widetilde{p}_i - p_i$ เอ็มพิริเคิลโลจิท(empirical logit) คือ $g(\widetilde{p}_i) = \log\{\widetilde{p}_i/(1-\widetilde{p}_i)\}$ ขณะที่เอ็มพิริเคิลโพรบิท(empirical probit) คือ $g(\widetilde{p}_i) = \Phi^{-1}(\widetilde{p}_i)$

$$g(p_i) = x_i'\boldsymbol{\beta} + \frac{\partial g(p_i)}{\partial p_i}(\tilde{p}_i - p_i)$$
$$= \eta_i + \frac{\partial \eta_i}{\partial p_i}(\tilde{p}_i - p_i) \qquad (2.36)$$

เมื่อ $\frac{\partial \eta_i}{\partial p_i}$ เป็นอนุพันธ์ของลิงค์ฟังก์ชันกับความสัมพันธ์ที่ สอดคล้องกับฟังก์ชันตอบสนองค่าเฉลี่ย (mean response function) สำหรับตัวแบบโลจิทและตัวแบบโพรบิทมีฟังก์ชันตอบสนองค่าเฉลี่ย (p_i) คือ $\Lambda(x_i'\beta)$และ $\Phi^{-1}(x_i'\beta)$ ตามลำดับ

ปัญหาของกำลังสองน้อยสุดแบบถ่วงน้ำหนัก(weighted least square : WLS) คือ ทำ

$$\sum_i \frac{(\tilde{p}_i - p_i)^2}{w_i}$$

ให้มีค่าน้อยลง ซึ่งสัมพันธ์กับ $\boldsymbol{\beta}$ เมื่อ w_i คือน้ำหนักที่ถ่วงด้วยการผกผันของความแปรปรวน วิธี FGLS คือ

$$b_{GLS} = [\mathbf{X}'\mathbf{W}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}g(\tilde{p}_i)$$

2.11.1 วิธีไคกำลังสองต่ำสุดแบบโลจิท (Minimum Logit Chi Square Method)

การประมาณไคกำลังสองต่ำสุดแบบโลจิทใช้ตัวแปรตามที่ได้จากเอ็มพิริเคิลโลจิท ตัวแบบเชิงเส้นสามารถเขียนได้ดังนี้

$$\text{logit}(\tilde{p}_i) = \log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = x_i'\beta + \varepsilon_i \qquad (2.37)$$

เมื่อ $E(\varepsilon) = 0$ และ $\text{var}(\varepsilon) = 1/[n_i p_i(1-p_i)]$ วิธีการประมาณนี้เป็นการประมาณที่ใช้ความน่าจะเป็นแบบเอ็มพิริเคิล (empirical probabilities) $\text{var}(\varepsilon) = 1/[n_i \tilde{p}_i(1-\tilde{p}_i)]$ ในกรณีนี้เราใช้ผลรวมกำลังสองแบบถ่วงน้ำหนัก(weighted sum of squares)ที่มีค่าน้อยสุด

$$\sum_i w_i[\log it(\tilde{p}_i) - x_i'\beta]^2$$

กับความสัมพันธ์ของ $\boldsymbol{\beta}$ ที่ใช้ FGLS ซึ่งมีการถ่วงน้ำหนักเท่ากับ $w_i = n_i\tilde{p}_i(1-\tilde{p}_i)$

2.11.2 วิธีไคกำลังสองต่ำสุดแบบโพรบิท (Minimum Probit Chi Square Method)

การประมาณไคกำลังสองต่ำสุดแบบโพรบิท ใช้ส่วนกลับของฟังก์ชันการแจกแจงสะสมแบบปกติ (หรือ Z -score) การตอบสนองที่นำไปสู่ ความน่าจะเป็นแบบเอ็มพิริเคิล (empirical probabilities) คือ $\Phi^{-1}(\widetilde{p}_i)$ ฟังก์ชันผกผันหาได้มากมายในโปรแกรมทางสถิติ ระหว่างฟังก์ชันการแจกแจงสะสมแบบปกติ (cumulative normal distribution function) ตัวแบบเชิงเส้นสามารถเขียนจากการใช้ เอ็มพิริเคิลโพรบิท ได้ดังนี้

$$\text{probit}(\widetilde{p}_i) = \Phi^{-1}(\widetilde{p}_i) = x_i'\beta + \varepsilon_i \tag{2.38}$$

เมื่อ $E(\varepsilon) = 0$ ความแปรปรวนของความคลาดเคลื่อน(error variance) คือ การประมาณ ดังนี้

$$\hat{\text{var}}(\varepsilon) = \frac{\widetilde{p}_i(1-\widetilde{p}_i)}{n_i \phi(\hat{z}_i)^2}$$

โดยที่ $\hat{z}_i = \Phi^{-1}(\widetilde{p}_i)$

เช่นเดียวกับตัวแบบโลจิท การประมาณค่ามีผลต่อผลรวมกำลังสองแบบถ่วงน้ำหนักที่น้อยสุดโดยใช้การผกผันของความแปรปรวนของความคลาดเคลื่อน(error variance)ของน้ำหนัก ดังนั้น การถดถอยแบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก(weighted least square :WLS) โดยใช้เอ็มพิริเคิลโพรบิท ซึ่งมีการถ่วงน้ำหนักเท่ากับ $w_i = n_i \phi(\hat{z}_i)^2 / \widetilde{p}_i(1-\widetilde{p}_i)$ เพื่อใช้ประมาณค่า β's

2.12 การทดสอบนัยสำคัญของค่าพารามิเตอร์

สมมติฐานการทดสอบคือ $H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0$

ค่าของการทดสอบสถิตินี้คือ ของ Wald Statistic ซึ่งกำหนดโดย

$$W = \frac{\hat{\beta}_j}{S.E.(\hat{\beta}_j)}$$

โดยที่ตัวสถิติ W มีการแจกแจงแบบปกติเมื่อตัวอย่างมีขนาดใหญ่ (Large–sample normal distribution) Ryan(1997) กล่าวคือ ค่าสถิติ W นี้ไม่ได้มีการแจกแจงแบบ t ถึงแม้ว่าจะมีรูปแบบการคำนวณที่คล้ายกับตัวสถิติ t ในการทดสอบนัยสำคัญของตัวแปรอิสระแต่ละตัวของการ

วิเคราะห์การถดถอยเชิงเส้นทั่วไป แต่อย่างไรก็ตามการใช้ Wald test ก็มีข้อเสียคือเมื่อค่าสัมบูรณ์ของค่าพารามิเตอร์มีค่ามาก ค่าคลาดเคลื่อนมาตรฐาน(standard error) ก็มักมีค่ามากด้วย ทำให้อัตราส่วน $\hat{\beta}/SE(\hat{\beta})$ หรือ Wald statistic มีค่าน้อย ซึ่งนำไปสู่ความผิดพลาดในการทดสอบ (reject null hypothesis) (กราบแก้ว, 2544 :31-35)

2.13 การทดสอบความเหมาะสมของตัวแบบการวิเคราะห์ (Goodness of Fit Test)

กำหนดสมมติฐานการทดสอบคือ

H_0 : ตัวแบบของการวิเคราะห์มีความเหมาะสม

H_1 : ตัวแบบของการวิเคราะห์ไม่มีความเหมาะสม

ค่าสถิติของการทดสอบนี้คือ Deviance ซึ่งกำหนดโดย

$$\begin{aligned} D &= -2\log\Lambda \\ &= -2\log\left(\frac{\hat{L}_c}{\hat{L}_f}\right) \\ &= -2\left(\log\hat{L}_c - \log\hat{L}_f\right) \end{aligned}$$

เมื่อ $\hat{L}_c$ เป็น maximized likelihood ของตัวแบบปัจจุบัน (current model) ซึ่งเป็นตัวแบบที่สร้างขึ้นจากข้อมูลตัวอย่าง และ $\hat{L}_f$ เป็น maximized likelihood ของตัวแบบเต็ม (full or saturated model) ซึ่งเป็นตัวแบบที่สมมติว่า ค่าประมาณที่ได้จากตัวแบบนั้นจะมีค่าเท่ากับข้อมูลจากค่าสังเกตตัวอย่าง หรือเป็นตัวแบบที่ fit กับข้อมูลได้ดีนั่นเอง

ถ้า $\hat{L}_c < \hat{L}_f$ แล้ว จะทำให้ Deviance มีค่ามาก แสดงว่าตัวแบบปัจจุบันที่สร้างขึ้นนั้นใช้เป็นตัวอธิบายข้อมูลได้ไม่ดี แต่ถ้า $\hat{L}_c = \hat{L}_f$ จะทำให้ Deviance มีค่าน้อยแสดงว่าตัวแบบปัจจุบันที่สร้างขึ้นนั้นไม่ต่างไปจากตัวแบบเต็ม หรือเป็นตัวแบบที่สามารถอธิบายข้อมูลได้ดีนั่นเอง ดังนั้นตัวสถิติ Deviance จึงเป็นตัวที่ใช้วัดว่าตัวแบบปัจจุบันนั้นมีความแตกต่างจากตัวแบบเต็มมากน้อยเพียงใด

ในตัวแบบที่ค่าสังเกตมีการแจกแจงแบบทวินาม (binomial observations) นั่นคือ ถ้าพิจารณาข้อมูลแบบเป็นกลุ่มกรณีที่ชุดของตัวแปรอธิบายมีค่าของข้อมูลเหมือนกัน โดยมีความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจในแต่ละกลุ่มของชุดตัวแปรอธิบายที่ i คือ $p_i = y_i/n_i, i = 1,2,\ldots,n$ เมื่อ y_i คือจำนวนของค่าสังเกต Y ที่มีค่าเท่ากับ 1 ในแต่ละกลุ่มของ

$p_i = y_i/n_i, i = 1,2,...,n$ เมื่อ y_i คือจำนวนของค่าสังเกต Y ที่มีค่าเท่ากับ 1 ในแต่ละกลุ่มของชุดในแต่ละตัวแปรอธิบายที่ i และ n_i คือจำนวนค่าสังเกต Y ทั้งหมดที่มีค่าตัวแปรอธิบายเหมือนกันในแต่ละชุดของตัวแปรอธิบายที่ i ซึ่งในกรณีนี้จะได้ฟังก์ชันภาวะน่าจะเป็น (likelihood function)

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{2.39}$$

จากการกำหนดตัวแบบถดถอยโลจิสติคที่มีพารามิเตอร์ $\beta_0, \beta_1, ..., \beta_k$ จำนวน $k+1$ ตัว จะได้ค่าความน่าจะเป็น (fitted probabilities) $\hat{p}_i$ คือ

$$\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_j x_{ji}$$

ดังนั้นจะได้ maximized log-likelihood function ของตัวแบบปัจจุบัน (current model) คือ

$$\log(\hat{L}_c) = \sum_{i=1}^{n} \left\{ \log\binom{n_i}{y_i} + y_i \log \hat{p}_i + (n_i - y_i)\log(1-\hat{p}_i) \right\} \tag{2.40}$$

ภายใต้ตัวแบบเต็ม (full model) ค่าประมาณความน่าจะเป็น จะเป็นค่าเดียวกันกับ สัดส่วนของค่าสังเกต $\tilde{p}_i = y_i/n_i, i = 1,2,...,n$ ดังนั้น ได้ maximized log-likelihood function ของตัวแบบเต็ม คือ

$$\log(\hat{L}_f) = \sum_{i=1}^{n} \left\{ \log\binom{n_i}{y_i} + y_i \log \tilde{p}_i + (n_i - y_i)\log(1-\tilde{p}_i) \right\}$$

ดังนั้น ตัวสถิติ Deviance จะมีค่าดังนี้

$$\begin{aligned} D &= -2(\log\hat{L}_c - \log\hat{L}_f) \\ &= 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\tilde{p}_i}{\hat{p}_i}\right) + (n_i - y_i)\log\left(\frac{1-\tilde{p}_i}{1-\hat{p}_i}\right) \right\} \end{aligned}$$

ถ้าค่าประมาณจำนวนของการเกิดเหตุการณ์ที่สนใจ (fitted number of success) ของตัวแบบปัจจุบันคือ $\hat{y}_i = n_i \hat{p}_i$ แล้ว ตัวสถิติ Deviance สามารถเขียนได้ในรูปของ

$$D = 2\sum_{i} \left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right\} \tag{2.41}$$

ซึ่งก็คือ ตัวสถิติที่เปรียบเทียบค่าสังเกต y_i ภายใต้ตัวแบบกับค่าประมาณ $\hat{y}_i$ ภายใต้ตัวแบบปัจจุบันนั่นเอง

อย่างไรก็ตาม Collett (2003) กล่าวว่า ตัวสถิติ Deviance จะไม่มีความหมายอะไร ถ้านำมาใช้ทดสอบความเหมาะสมของตัวแบบที่ค่าสังเกตมีการแจกแจงแบบเบอร์นูลลี (Bernoulli observations) ซึ่งในที่นี้ $n_i = 1$ และฟังก์ชันภาวะน่าจะเป็น คือ

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

จะได้ maximized log-likelihood ภายใต้ตัวแบบปัจจุบันดังนี้

$$\log(\hat{L}_c) = \sum_{i=1}^{n} \{y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\}$$

สำหรับตัวแบบเต็มแล้ว ค่าประมาณความน่าจะเป็น (fitted probabilities) จะเป็นค่าเดียวกันกับค่าสังเกตของ y ดังนั้น $\hat{p}_i = y_i$ และเนื่องจากว่า $y_i \log y_i$ และ $(1 - y_i)\log(1 - y_i)$ มีค่าเป็น 0 สำหรับค่าที่เป็นไปได้ 2 ค่า ของ y_i คือ 0 และ 1 ซึ่งทำให้ได้ $\log(\hat{L}_c) = 0$ ดังนั้น ตัวสถิติ Deviance ในกรณีนี้คือ

$$\begin{aligned} D &= -2\sum_{i=1}^{n} \{y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\} \\ &= 2\sum_{i=1}^{n} \left\{y_i \log\left(\frac{\hat{p}_i}{(1-\hat{p}_i)}\right) + \log(1 - \hat{p})\right\} \end{aligned} \tag{2.42}$$

ถ้าทำการหาอนุพันธ์ของ $\log L(\beta)$ เทียบกับ β จะได้

$$\begin{aligned} \frac{\partial \log L(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \left\{\frac{y_i}{p_i} - \left(\frac{1-y_i}{1-p_i}\right)\right\} p_i(1 - p_i)x_{ji} \\ \sum_{j=1}^{k} \beta_j \frac{\partial \log L(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} (y_i - p_i) \sum_{j=1}^{k} \beta_j x_{ji} \\ &= \sum_{i=1}^{n} (y_i - p_i) \log\{p_i/(1 - p_i)\} \end{aligned}$$

เนื่องจาก $\hat{\beta}$ เป็น maximized likelihood estimator ของ β จึงทำให้อนุพันธ์ทางซ้ายของสมการมีค่าเป็น 0 ที่ $\hat{\beta}$ ทำให้ได้ $\sum_{i=1}^{n}(y_i - \hat{p}_i)\,\mathrm{logit}\{\hat{p}_i\} = 0$

ดังนั้น
$$\sum_{i=1}^{n} y_i \ \mathrm{logit}(\hat{p}_i) = \sum_{i=1}^{n} \hat{p}_i \ \mathrm{logit}(\hat{p}_i)$$

ซึ่งถ้าแทนเทอมของ $\sum_{i=1}^{n} y_i \ \mathrm{logit}(\hat{p}_i)$ ลงในสมการ (2.42) จะได้ค่าสถิติ Deviance คือ

$$D = -2\sum_{i=1}^{n}\{\hat{p}_i \ \mathrm{logit}(\hat{p}_i) + \log(1-\hat{p}_i)\} \tag{2.43}$$

ซึ่งจะเห็นว่าค่า Deviance ในสมการ (2.43) นี้ไม่ได้ให้การเปรียบเทียบอะไรระหว่างค่าสังเกตของ $\hat{p}_i$ ภายใต้ตัวแบบปัจจุบัน ในกรณีที่ค่าสังเกตมีการแจกแจงแบบเบอร์นูลลี

ตัวสถิติ Deviance นี้มีการแจกแจงโดยประมาณแบบไคกำลังสองเมื่อตัวอย่างเข้าใกล้อนันต์ โดยมีองศาความเป็นอิสระเท่ากับ $n - p$ เมื่อ n คือ จำนวนของค่าสังเกตแบบทวินาม และ p คือ จำนวนของพารามิเตอร์ที่ไม่ทราบค่าในตัวแบบย่อย

2.14 การศึกษางานวิจัยที่เกี่ยวข้อง

ผู้ช่วยศาสตราจารย์ ดร. ทิติยา จิตติหรรษา ได้ทำการศึกษาการเปรียบเทียบสารสกัดหยาบจากพืชบางชนิดที่มีผลต่อการตายของเพลี้ยจั๊กจั่นสีเขียว โดยใช้ตัวแบบโพรบิทช่วยในการวิเคราะห์ผลปรากฏว่าสารสกัดหยาบที่มีผลต่อการตายของเพลี้ยจั๊กจั่นสีเขียวคือผลมะระขี้นกและรากหนอนตายหยาก

กาญจนา พานิชการ(2539) ศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ ในสมการถดถอยโลจิสติค ซึ่งวิธีการประมาณค่าพารามิเตอร์ที่ใช้ศึกษา คือ วิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันจำแนกประเภท และวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลที่ไม่ได้จัดกลุ่ม ตัวแปรตาม Y มี 2 ค่า คือ 0 หรือ 1 โดยทำการเปรียบเทียบในกรณีที่ตัวแปรอธิบายมีการแจกแจง 3 ลักษณะ คือ การแจกแจงแบบปกติ การแจกแจงแบบชี้กำลัง และการแจกแจงแบบไวบูลล์ ซึ่งเกณฑ์ที่ใช้เปรียบเทียบคือ ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย และตัวสถิติ Deviance ผลสรุปที่ได้คือวิธีภาวะน่าจะเป็นสูงสุดให้ค่า RMSE น้อยกว่าวิธีฟังก์ชันจำแนกประเภท ยกเว้นกรณีที่มีค่าสัดส่วนสูง $(P > 0.75)$ และมีขนาดตัวอย่างเล็ก $(n < 30)$

ชนิศวรา ฉัตรแก้ว (2543) ได้ศึกษาเปรียบเทียบการวิเคราะห์การถดถอยเมื่อตัวแปรตามมีค่าเป็น 2 ลักษณะ กรณีข้อมูลเฉพาะบุคคลโยใช้ตัวแบบความน่าจะเป็นเชิงเส้น ตัวแบบความน่าจะเป็นเชิงเส้นแบบถ่วงน้ำหนัก ตัวแบบโพรบิท และตัวแบบโลจิท ได้ผลการศึกษาคือ ตัวแบบโพรบิทและตัวแบบโลจิทมีความเหมาะสมมากกว่าตัวแบบความน่าจะเป็นเชิงเส้นและตัวแบบความน่าจะเป็นเชิงเส้นแบบถ่วงน้ำหนัก และยังพบว่าตัวแบบโพรบิทมีความเหมาะสมมากกว่าตัวแบบโลจิท เพราะให้ค่า Pseudo-R^2 ถูกต้องมากกว่าตัวแบบโลจิท

ทัศนาพร จงเกตุกรณ์ (2547) ศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบโลจีสติคทวินาม ด้วยวิธีภาวะน่าจะเป็นสูงสุด(MLE) วิธีการถ่วงน้ำหนัก(WE) และวิธีปรับแก้เบื้องต้น (PC) ผลปรากฏว่า เมื่อค่าเฉลี่ยความน่าจะเป็นของเหตุการณ์ที่สนใจของประชากรเท่ากับ 0.1 ,0.3 ด้วยวิธี MLE ให้ค่าระยะห่างมาหาโลบิสเฉลี่ย(AMH) น้อยสุด แต่ในกรณีที่ค่าเฉลี่ยความน่าจะเป็นของเหตุการณ์ที่สนใจของประชากรเท่ากับ 0.5 ,0.8 ด้วยวิธี PC ให้ค่า AMH น้อยสุด

เรวดี เรืองอยู่ (2547) ศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยโลจิสติค วิธีการประมาณค่าพารามิเตอร์ที่ใช้ในการศึกษาครั้งนี้ คือ วิธีการแบบริดจ์(RE) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และวิธีฟังก์ชันจำแนกประเภท(DF) โดยที่ตัวแปรตาม Y เป็นตัวแปรเชิงคุณภาพมี 2 ค่า คือ 0 หรือ 1 และมีตัวแปรอธิบาย X 1 หรือ 2 ตัวแปร การเปรียบเทียบกระทำภายใต้ขนาดตัวอย่าง $n = 10,20,30,40$ โดยมีสัดส่วนของตัวแปรตาม $(Y = 1)$ ค่อนข้างสูง $(P \geq 0.75)$ คือ 0.75, 0.80, 0.85, 0.90 และ 0.95 และการแจกแจงของตัวแปรอธิบาย คือ การแจกแจงแบบปกติ การแจกแจงแบบชี้กำลัง และการแจกแจงแบบไวบูลล์ ผลสรุปกรณีตัวแปรอธิบาย 1 ตัว วิธี MLE ให้ค่า RMSE และค่า DV น้อยกว่าวิธี DF ยกเว้นเมื่อขนาดตัวอย่าง n = 30 และ 40 วิธี DF ให้ค่า RMSE และค่า DV น้อยกว่าวิธี MLE กรณีตัวแปรอธิบาย 2 ตัว ที่มีการแจกแจงแบบปกติและการแจกแจงแบบชี้กำลัง หรือ การแจกแจงแบบปกติและการแจกแจงแบบไวบูลล์ หรือการแจกแจงแบบชี้กำลังและการแจกแจงแบบไวบูลล์ วิธี MLE ให้ค่า RMSE และ DV น้อยกว่าวิธีอื่น

Takeshi Amemiya (1974) ได้ศึกษาวิธีการประมาณค่าไคกำลังสองต่ำสุดแบบโพรบิทกับข้อมูลจริงทำการเปรียบเทียบกับวิธี MLE ทำการศึกษาเมื่อตัวแปรตอบสนองเป็นเป็นตัวแปรเชิงสุ่มแบบสองกลุ่ม โดยตัวแปรตอบสนองมีสองตัว คือ การตอบสนองของผู้ป่วยที่ขาดลมหายใจแทนด้วย Y และการตอบสนองของผู้มีอาการหอบแทน ด้วย Z และใช้ตัวแบบโพรบิทในการวิเคราะห์ข้อมูล โดยทำการวิเคราะห์ตัวแปรตอบสนองทีละตัว และทำการวิเคราะห์เมื่อตัวแปรตอบสนองมีสองตัวตามลำดับ โดยมีตัวแปรอธิบายคือค่ากลางของ ตั้งแต่ 20-64 แบ่งเป็น 9 กลุ่ม และการหาความสัมพันธ์ตัวแปรตอบสนองสองตัวคือ ρ_{YZ} ผลปรากฏว่าวิธี ML ให้ค่า ρ_{YZ} เท่ากับ 0.7709 ส่วนวิธี FIMCS ให้ค่า ρ_{YZ} เท่ากับ 0.7746 ต่อมาปี (1976) ได้ทำการศึกษาเกี่ยวกับตัวแบบการตอบสนองคุณภาพทั่วไปรวมถึงตัวแปรหลายประเภทต่างๆ โดยเฉพาะการศึกษาวิธีการ

ประมาณค่าพารามิเตอร์ด้วยไคกำลังสองต่ำสุดสำหรับข้อมูลการตอบสนองแบบคุณภาพทั่วไป นอกจากวิธีการประมาณแบบ MCS แล้ว ยังศึกษาถึงวิธี MLE และ WLS อีกด้วย

Jonathan Nagler (1994:230-255) ได้ทำการศึกษาของการเปลี่ยนแปลงของตัวประมาณค่าที่นำไปสู่ตัวแบบโพรบิท (probit model) และตัวแบบโลจิท (logit) โดยการทดลองให้เห็นถึงการลงคะแนนเสียง กับจุดเริ่มต้นของความน่าจะเป็นของการลงคะแนนที่น้อยกว่า 0.5 ที่เปลี่ยนในตัวแปรอธิบาย และได้ทำการสำรวจกับบุคคลที่มีการศึกษาต่ำ หรือบุคคลที่มีการศึกษาสูง มีสิทธิในการเปลี่ยนแปลงตามกฎการลงคะแนน กับความเกี่ยวข้องของความน่าจะเป็นในการลงคะแนน

J.Econ.Entomol. (1995:1513-1516) ได้ทำการศึกษาการเปลี่ยนแปลงของ คอมพลีเมนทารีล็อก-ล็อก (complementary log-log) โลจิท (logit) โพรบิท (probit) ล็อก-คอมพลีเมนทารีล็อก-ล็อก (log complementary log-log) ล็อกโลจิท (log logit) และการแปลงโพรบิท (probit transformation) ข้อมูลจากประสบการณ์ด้านbioassay กลับนำไปสู่หน่วยดั้งเดิมของการวัดผลสัดส่วนของการทดสอบเริ่มต้นต่อการตอบสนอง เมื่อทดสอบกับสิ่งกระตุ้น และคำนวณส่วนเหลือและส่วนเหลือมาตรฐาน แสดงถึงวิธีการที่สามารถใช้เลือกตัวแบบที่ดีที่สุดให้เหมาะสมกับข้อมูล ต่อมาในปีเดียวกันใช้วิธีการคำนวณทางสถิติสำหรับตารางมรณะ โดยใช้การถดถอยคอมพลีเมนทารีล็อก-ล็อก , โลจิท, การแปลงโพรบิท ของสัดส่วนของตัวแปรตอบสนองที่ไม่ได้เปลี่ยนแปลงตามเวลา หรือการแปลง logarithmic ของเวลา และการคำนวณภายใต้เงื่อนไขลิมิต

Agresti (1990) ซึ่งนำข้อมูลการตายของแมลงปีกแข็งเมื่อได้รับปริมาณสารพิษ(log-dose)จากงานทดลองของ Bliss (1935) ทำการเปรียบเทียบตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก โดยมีวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด(maximum likelihood estimation method : MLE) ผลการปรากฏว่า ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ให้ค่า Deviance ต่ำสุด รองลงมา คือ ตัวแบบโพรบิท และ ตัวแบบโลจิท นั้น หมายความว่าตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เป็นตัวแบบที่ดีที่สุด สำหรับข้อมูลการตายของแมลงปีกแข็ง

Berkson (1955) ทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ด้วย MLE และ MCS ของตัวแบบโพรบิท ผลปรากฏว่าวิธี MCS ให้ค่าความแปรปรวนน้อยกว่าวิธี MLE และ Berkson (1955) ได้แนะนำว่าควรใช้วิธี MCS กับข้อมูลจริงทางชีววิทยา เพราะคำนวณง่ายไม่ยุ่งยาก

Faqir Muhammad และคณะ(1990) ได้ทำการวิเคราะห์การถดถอยโลจิสติคในการตอบสนองของสารพิษ โดยการประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธี OLS, MLE,WLS ผลสรุปว่าในการวิเคราะห์การถดถอยโลจิสติคผลจากการประมาณค่าพารามิเตอร์ด้วยวิธี WLS ให้ผลดีกับข้อมูลของการทดลองทางชีววิทยานี้

Huhm,M (2000) ได้ทำการศึกษาการเปรียบเทียบวิธี MLE , MCS และ LS ซึ่งใช้กับข้อมูลจริงเกี่ยวกับหัวบีตที่ใช้ทำน้ำตาล ผลการทดลอง (z1,z2, และ z3) และผลของการประมาณที่สามารถหาได้โดยวิธีภาวะน่าจะเป็นสูงสุด $\left(\hat{R}_1\right)$ วีธีไคกำลังสองต่ำสุด $\left(\hat{R}_2\right)$ และวิธีกำลังสองน้อย

สุด $(\hat{R}_3)$ ผลปรากฏจากตารางสรุปได้ว่าวิธีไคกำลังสองต่ำสุด $(\hat{R}_2)$ มีค่ามากกว่าเท่ากับ หรือน้อยกว่าเท่ากับ การประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด $(\hat{R}_1)$ ผลที่ได้เป็นไปตามทฤษฎี การประมาณทั้งสองวิธีคือ $(\hat{R}_1)$และ $(\hat{R}_2)$ ให้ค่าใกล้เคียงกันมาก และสอดคล้องกับทฤษฎี ในขณะที่วิธี LS ให้การประมาณค่าต่างจากพวก

Cramer (2003) ได้กล่าวว่าตัวแบบโลจิทคำนวณได้ง่ายกว่า และอาจเป็นเหตุผลที่มีผู้ใช้ตัวแบบโลจิทเพิ่มขึ้นและมากกว่าการใช้ตัวแบบโพรบิท เช่น ในปี 1990-1994 มีงานที่ใช้ตัวแบบโลจิท 311 ชิ้น ส่วนตัวแบบโพรบิทมีปรากฏ 127 ชิ้น

Teijin Twaron (2006) ได้ทำการศึกษาการพยากรณ์ของตัวแบบทั้ง 3 ตัวแบบโดยการจำลองข้อมูลที่เกี่ยวกับอัตราความเร็วของกระสุนปืน (X) และล็อกของอัตราความเร็ว $\log(X)$ ที่มีผลต่อการยิงทะลุเสื้อเกาะ โดยใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood estimation method :MLE) ผลปรากฏว่าตัวแบบโลจิท ของ (X) และ $\log(X)$ ให้ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองต่ำสุด (mean square error : MSE) ดังนั้นตัวแบบโลจิท จึงเป็นตัวแบบที่เหมาะสมกับข้อมูลอัตราเร็วของกระสุนปืน

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินงานวิจัยนี้เป็นการวิจัยเพื่อการเปรียบเทียบในการวิเคราะห์ด้วย ตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก โดยนำวิธีการของทั้ง 3 ตัวแบบมาประยุกต์กับข้อมูลจริงที่เราหามาได้จากหลากหลายสาขาวิชา เช่น ด้านการแพทย์ วิทยาศาสตร์ วิศวกรรมศาสตร์ สังคม และอื่นๆ โดยเปรียบเทียบว่าตัวแบบใดเหมาะสมกับลักษณะของข้อมูล เมื่อวิธีการประมาณค่าพารามิเตอร์ของตัวแบบต่างกัน โดยใช้วิธีการประมาณค่าพารามิเตอร์ 3 วิธีคือ วิธีกำลังสองน้อยสุดถ่วงน้ำหนัก (weighted least squares method :WLS) วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood estimation method :MLE) หรือ วิธีไคกำลังสองต่ำสุด (minimum chi-square method : MCS) และสามารถใช้ตัวแบบพยากรณ์เหตุการณ์ในอนาคตได้ โดยมีแผนการดำเนินงานวิจัยดังนี้

3.1 แผนการดำเนินการวิจัย

ในการดำเนินการวิจัยครั้งนี้ ทำการหาข้อมูล ทางด้านการแพทย์ ด้านวิทยาศาสตร์ วิศวกรรมศาสตร์ และ ทางสังคมศาสตร์ ที่ใช้ในการวิเคราะห์ สำหรับศึกษาและเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก วิธีภาวะน่าจะเป็นสูงสุด และวิธีไคกำลังสองต่ำสุด ดังนี้

1. ใช้ข้อมูลจริงทางการแพทย์ 3 ชุด
2. ใช้ข้อมูลจริงทางด้านวิทยาศาสตร์(ชีววิทยา) 3 ชุด
3. ใช้ข้อมูลจริงด้านวิศวกรรม 1 ชุด
4. ใช้ข้อมูลจริงด้านสังคม(การเลือกตั้ง) 1 ชุด
5. ใช้ข้อมูลจริงด้านการศึกษา 1 ชุด
6. ตัวแปรอธิบาย(X) ที่ใช้ในการวิจัยของแต่ละชุดข้อมูลมีเพียงตัวแปรเดียว
7. ตัวแปรอธิบายเป็นข้อมูลเชิงกลุ่มหรือข้อมูลเชิงประมาณก็ได้
8. ตัวแปรตอบสนอง (Y_i)เป็นอิสระซึ่งกันและกันและมีการแจกแจงแบบทวินาม
9. ขนาดตัวอย่างที่ใช้ในงานวิจัยขึ้นอยู่กับลักษณะข้อมูลที่ใช้ในการวิเคราะห์
10. การวิจัยครั้งนี้ได้ทำการประมาณค่าพารามิเตอร์ของแต่ละตัวแบบด้วยโปรแกรม R

3.2 ขั้นตอนในการดำเนินงานวิจัย

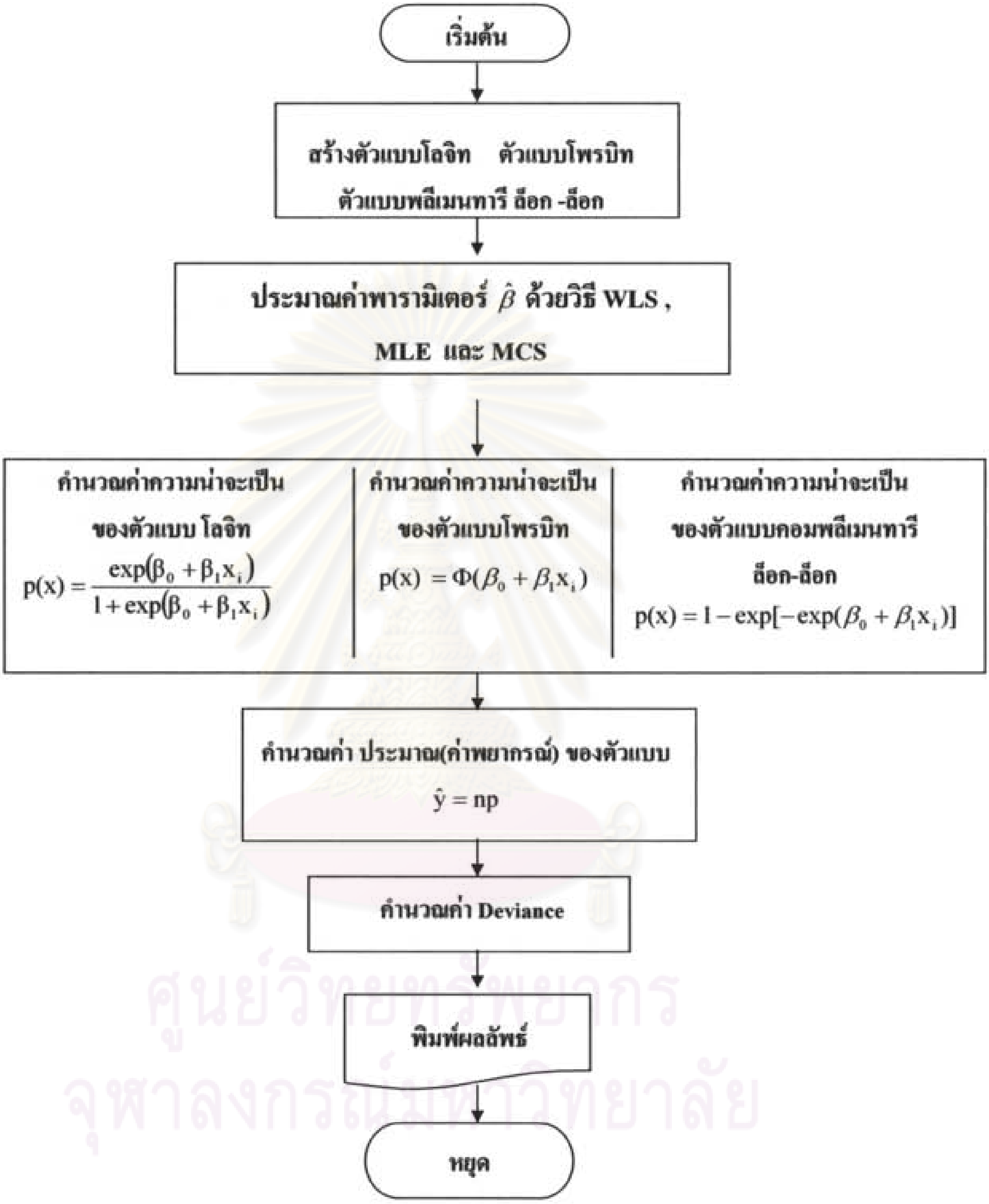
ขั้นตอนในการดำเนินงานวิจัยมีดังนี้ คือ

1. ศึกษาตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบตัวแบบคอมพลีเมนทารี ล็อก-ล็อก
2. ศึกษาวิธีการประมาณค่าพารามิเตอร์ 3 วิธี คือวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method :WLS วิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE) หรือ วิธีไคกำลังสองต่ำสุด (Minimum Chi-Square method : MCS)
3. ศึกษาวิธีการเขียนโปรแกรมเพื่อทำการเขียนตามสถานการณ์ของงานทดลองแต่ละสาขาที่ใช้ในการวิจัย
4. สร้างข้อมูลตามสถานการณ์ของงานทดลองแต่ละสาขาที่ใช้ในการวิจัย
    4.1 ข้อมูลทางด้านการแพทย์ของ Draper,Voller and Carpenter(1972)
    4.2 ข้อมูลทางด้านการแพทย์ของ Ashford and Sowden(1970)
    4.3 ข้อมูลทางด้านการแพทย์ของ Cornfield (1962)
    4.4 ข้อมูลทางด้านชีววิทยา Martin(1942)
    4.5 ข้อมูลทางด้านชีววิทยา Muhammad(1990)
    4.6 ข้อมูลทางด้านชีววิทยา Strand(1930)
    4.7 ข้อมูลทางด้านการวิศวกรรม Montgomery and Peck(1982)
    4.8 ข้อมูลทางด้านสังคมศาสตร์ Shockey (1988)
    4.9 ข้อมูลทางด้านการศึกษา Haberman(1978)
5. เขียนโปรแกรมสำหรับคำนวณการสร้างตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ด้วยวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares method :WLS) เพื่อหาค่าพารามิเตอร์ β_0,β_1 และหลังจากได้ค่าพารามิเตอร์แล้วทำการเขียนโปรแกรมสำหรับการคำนวณหาความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ และทำการเขียนโปรแกรมสำหรับการพล็อตกราฟความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก
6. เขียนโปรแกรมสำหรับคำนวณการสร้างตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ด้วยวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด(Maximum Likelihood Estimation method :MLE) เพื่อหาค่าพารามิเตอร์ β_0,β_1 และหลังจากได้ค่าพารามิเตอร์แล้วทำการเขียนโปรแกรมสำหรับการ

คำนวณหาความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ และทำการเขียนโปรแกรมสำหรับการพล็อตกราฟความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีภาวะน่าจะเป็นสูงสุด

7. เขียนโปรแกรมสำหรับคำนวณการสร้างตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ด้วยวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีไคกำลังสองต่ำสุด (Minimum Chi-Square method : MCS) เพื่อหาค่าพารามิเตอร์ β_0, β_1 และหลังจากได้ค่าพารามิเตอร์แล้วทำการเขียนโปรแกรมสำหรับการคำนวณหาความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ และทำการเขียนโปรแกรมสำหรับการพล็อตกราฟความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีไคกำลังสองน้อยสุด
8. เขียนโปรแกรมสำหรับคำนวณค่าประมาณของตัวแบบโลจิท ด้วยวิธีประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก วิธีภาวะน่าจะเป็นสูงสุด และวิธีไคกำลังสองต่ำสุด
9. เขียนโปรแกรมสำหรับคำนวณค่าประมาณของตัวแบบโพรบิท ด้วยวิธีประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก วิธีภาวะน่าจะเป็นสูงสุด และวิธีไคกำลังสองต่ำสุด
10. เขียนโปรแกรมสำหรับคำนวณค่าประมาณของตัวแบบคอมพลีเมนทารีล็อก-ล็อก ด้วยวิธีประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก วิธีภาวะน่าจะเป็นสูงสุด และวิธีไคกำลังสองต่ำสุด
11. เขียนโปรแกรมสำหรับการพล็อตกราฟความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี
12. เขียนโปรแกรมสำหรับคำนวณค่าสถิติ Deviance ในการเปรียบเทียบตัวแบบ 3 ตัวแบบ และวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี
13. เขียนโปรแกรมสำหรับคำนวณหาค่าไคกำลังสอง (χ^2_{n-p}) หลังจากทราบค่าตัวสถิติ Deviance เพื่อตรวจสอบความเหมาะสมของตัวแบบ
14. ทำการวิเคราะห์ข้อมูลที่ได้จากการทดลองของแต่ละสาขาวิชา และนำผลการวิเคราะห์มาเปรียบเทียบแต่ละตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ ด้วยตัวสถิติ Deviance
15. ทำการเลือกตัวแบบ และ วิธีการประมาณค่าพารามิเตอร์ให้เหมาะสมกับลักษณะของข้อมูลที่ทำการวิจัยและสรุปผล

ผังการแสดงขั้นตอนในการวิจัย



3.2.1 สร้างข้อมูลที่ใช้ในการวิจัย

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

2. สร้างตัวแบบโพรบิท $\Phi^{-1}(P(x)) = \beta_0 + \beta_1 x$

3. สร้างตัวแบบคอมพลีเมนทารีล็อก-ล็อก $\log[-\log\{1-P(x)\}] = \beta_0 + \beta_1 x$

การเลือกลิงค์ฟังก์ชันสามารถหาได้โดยใช้ฟังก์ชันเชื่อม(link function) คือ

1. โลจิทหรือ โลจิสติคฟังก์ชัน

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

2. โพรบิทฟังก์ชัน

$$g(p) = \Phi^{-1}(p)$$

3. คอมพลีเมนทารีล็อก-ล็อก ฟังก์ชัน

$$g(p) = \log[-\log\{1-p\}]$$

3.2.2 หาค่าประมาณพารามิเตอร์

มีขั้นตอนดังนี้

3.2.2.1 วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก WLS

วิธีกำลังสองน้ำสุดแบบถ่วงน้ำหนัก ใช้ $V = 1/\mathbf{W}$ เมื่อ $\mathbf{W}$ แทน Variance-Covariance matrix ของ Y_i's ซึ่ง $V_i = \frac{1}{[\mathrm{Var}(Y_i)]}$

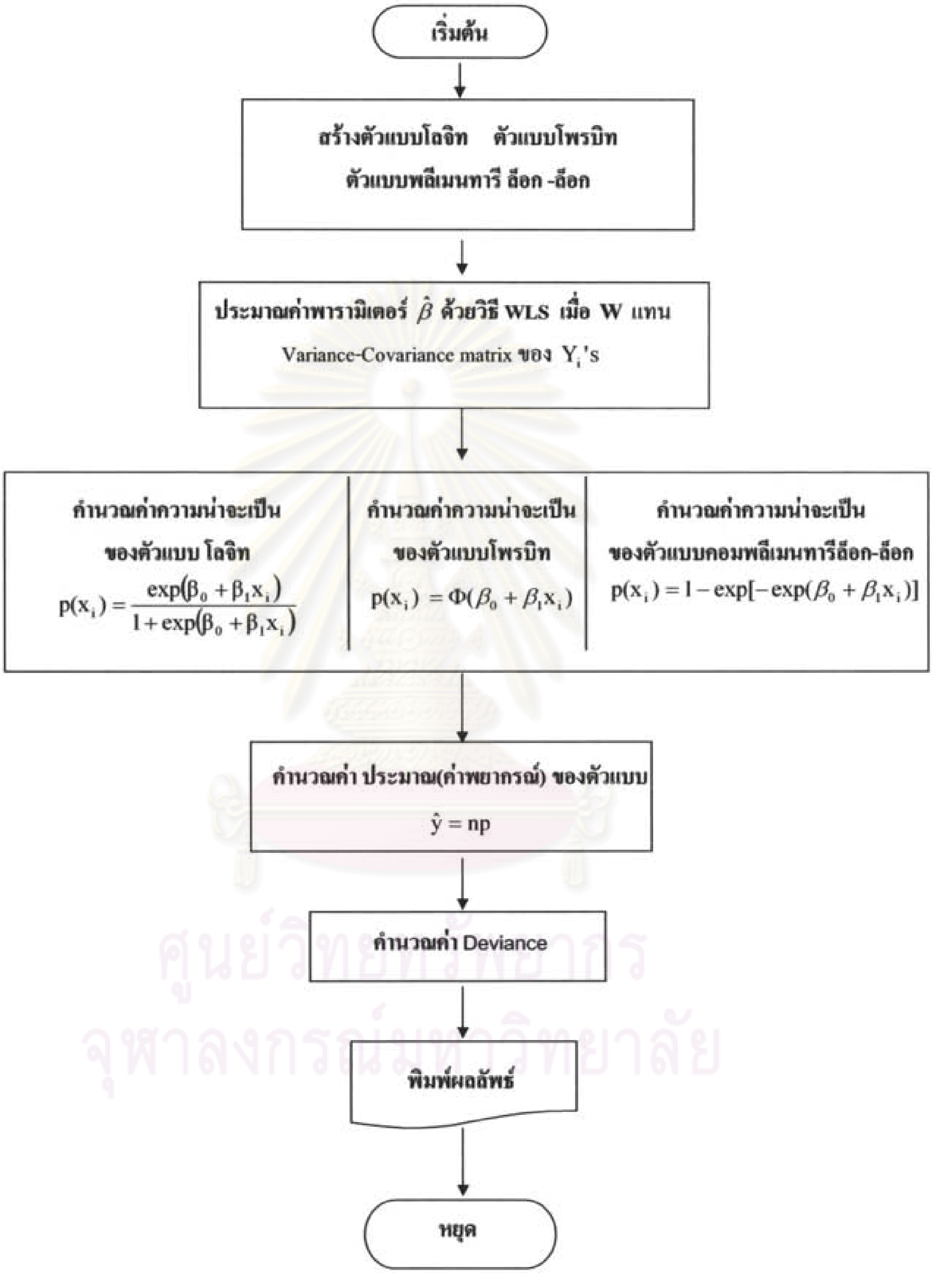
ดังนั้นค่าประมาณ $\hat{\boldsymbol{\beta}}$ จากวิธี **WLS** สามารถคำนวณจากสมการปกติ

$$\begin{aligned} \mathbf{X'W^{-1}(Y - \boldsymbol{\beta} X)} &= \mathbf{0} \\ \mathbf{X'W^{-1}X\boldsymbol{\beta}} &= \mathbf{X'W^{-1}Y} \end{aligned}$$

นั่นคือ $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'W^{-1}X}\right)^{-1}\mathbf{X'W^{-1}Y}$

โดยที่ Y สร้างตามลิงค์ฟังก์ชันของแต่ละตัวแบบ

วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก



3.2.2.2 วิธีภาวะน่าจะเป็นสูงสุด MLE

พิจารณาตัวประมาณค่าพารามิเตอร์ที่มีฟังก์ชันภาวะน่าจะเป็น ซึ่งจะใช้ Fisher scoring หรือใช้วิธี Newton –Raphson โดยทำการหาอนุพันธ์อันดับที่หนึ่งของลอการึม ของฟังก์ชันภาวะน่าจะเป็น $L(\beta)$ เทียบกับพารามิเตอร์ที่ละตัว ซึ่งคือ U(B) และหา อนุพันธ์อันดับที่สองของ $L(\beta)$ ซึ่งคือ H(B) จากนั้นนำค่า U(B)และ H(B) มาแทนลง ในสมการ $\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{H}^{(s)}\right)^{-1}\mathbf{U}^{(s)}$ จะมีการทำซ้ำจนกว่าจะได้ค่า $\hat{\boldsymbol{\beta}}$ ที่มีค่าไม่ต่างกัน มากเป็นที่ยอมรับได้

ดังนั้นตัวประมาณภาวะน่าจะเป็นสูงสุด คือ ลิมิตของ $\boldsymbol{\beta}^{(s)}$ ในขณะที่ $s \to \infty$ กล่าวอีกนัยหนึ่ง คือ ตัวประมาณภาวะน่าจะเป็นของตัวแบบเชิงเส้นที่วางนัยทั่วไป เป็นผล ของการย้อนซ้ำที่ใช้การถ่วงน้ำหนักของกำลังสองน้อยที่สุด โดยเมทริกซ์ของการถ่วง น้ำหนักจะเปลี่ยนค่าไปทุกๆครั้งของการย้อนซ้ำ เรียกกระบวนการนี้ว่า **"IRLS process"** (คือ Iterative Reweighted Least Square process) กระบวนการย้อนซ้ำทั้งหมดจะเสร็จสิ้น ต่อเมื่อค่าประมาณของ β ที่มีการย้อนซ้ำสองครั้งติดต่อกัน ให้ค่าผลต่างที่เล็กเพียงพอ หรือ เป็นศูนย์

อนึ่งความแปรปรวนร่วมเมื่อใกล้อนันต์ของ $\hat{\boldsymbol{\beta}}$ อยู่ในรูปเมทริกซ์ ที่เป็นส่วนกลับ ของเมทริกซ์ **M** เราสามารถหาค่าประมาณของ $Cov(\boldsymbol{\beta})$ จาก $C\hat{o}v(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$ โดยที่ $\hat{\mathbf{W}}$ คำนวณจากค่า **W** ณ $\hat{\boldsymbol{\beta}}$ เช่น $w_i = \left(\partial\mu_i / \partial\eta_i\right)^2 \;/ \mathrm{Var}(Y_i)$

โดยสรุปตัวแบบเชิงเส้นที่วางนัยทั่วไป ในการประมาณค่าพารามิเตอร์ของตัวแบบ ไม่ว่าจะใช้วิธี Fisher scoring หรือใช้วิธี Newton –Raphson โดยตรง ทั้งสองวิธีนี้พบว่ามีสิ่ง ที่ร่วมกัน คือ กระบวนการย้อนซ้ำต่างก็อาศัยวิธีภาวะน่าจะเป็นสูงสุดก่อน แล้วใช้การ ประมาณค่าจากสมการปกติในลักษณะของวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก โดยมี กระบวนการสร้างตัวถ่วงน้ำหนักขึ้นใหม่ในทุกๆรอบของการย้อนซ้ำ จนกระทั่งได้ตัว ประมาณที่ลู่เข้าและมีคุณสมบัติตามต้องการ เช่น ความพอเพียง ความคงเส้นคงวา ฯลฯ วิธี ดังกล่าวนี้อาจเรียกรวมกันได้ว่า การประมาณค่าภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ หรือวิธี IRLS ก็ได้ โดยเป็นวิธีประมาณค่าที่นิยมใช้สำหรับตัวแบบเชิงเส้นที่วางนัยทั่วไป ซึ่ง หมายรวมถึงตัวแบบการถดถอยโลจิสติค ตัวแบบโลจิท ตัวแบบล็อกลิเนียร์ ตัวแบบอื่นๆ ในกลุ่มเอ็กโปเนนเชียล และตัวแบบที่ไม่เชิงเส้นทั่วไปด้วย

วิธีภาวะน่าจะเป็นสูงสุด

เริ่มต้น

↓

สร้างตัวแบบโลจิท ตัวแบบโพรบิท
ตัวแบบพลีเมนทารี ล็อก -ล็อก

↓

ประมาณค่าพารามิเตอร์ $\hat{\beta}$ ด้วยวิธี MLE

↓

คำนวณค่าความน่าจะเป็นของตัวแบบ โลจิท	คำนวณค่าความน่าจะเป็นของตัวแบบโพรบิท	คำนวณค่าความน่าจะเป็นของตัวแบบคอมพลีเมนทารีล็อก-ล็อก
$p(x_i) = \dfrac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$	$p(x_i) = \Phi(\beta_0 + \beta_1 x_i)$	$p(x_i) = 1 - \exp[-\exp(\beta_0 + \beta_1 x_i)]$

↓

คำนวณค่า ประมาณ(ค่าพยากรณ์) ของตัวแบบ

$\hat{y} = np$

↓

คำนวณค่า Deviance

↓

พิมพ์ผลลัพธ์

↓

หยุด

3.2.2.3 วิธีไคกำลังสองต่ำสุด MCS

กำการสร้างตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก โดยมีการใช้ค่าถ่วงน้ำหนัก W ของแต่ละตัวแบบดังนี้

1. ค่าถ่วงน้ำหนักของตัวแบบโลจิท $w_i = n_i \tilde{p}_i (1 - \tilde{p}_i)$
2. ค่าถ่วงน้ำหนักของตัวแบบโพรบิท $w_i = n_i \phi(\hat{z}_i)^2 / \tilde{p}_i (1 - \tilde{p}_i)$
3. ค่าถ่วงน้ำหนักของตัวแบบคอมพลีเมนทารีล็อก-ล็อก

$$w_i = \{(\partial \hat{\eta}_i / \partial \hat{\mu}_i)^2 n_i \tilde{p}_i (1 - \tilde{p}_i)\}^{-1},$$
$$\partial \hat{\eta}_i / \partial \hat{\mu}_i = \{-\log(1 - \tilde{p}_i) n_i (1 - \tilde{p}_i)\}^{-1}$$

คำนวณหาค่าพารามิเตอร์ $\hat{\beta}$ จากวิธี MCS สามารถคำนวณจากสมการปกติ

$$b_{GLS} = [\mathbf{X'WX}]^{-1} \mathbf{X'W} g(\tilde{p}_i)$$

โดยที่ $\tilde{p}_i = y_i / n_i$ และ $\varepsilon_i = \tilde{p}_i - p_i$ เอ็มพิริเคิลโลจิท(empirical logit) คือ $g(\tilde{p}_i) = \log\{\tilde{p}_i / (1 - \tilde{p}_i)\}$ ขณะที่เอ็มพิริเคิลโพรบิท(empirical probit) คือ $g(\tilde{p}_i) = \Phi^{-1}(\tilde{p}_i)$ และ เอ็มพิริเคิลคอมพลีเมนทารี ล็อก-ล็อก คือ $g(\tilde{p}_i) = \log[-\log\{1 - \tilde{p}\}]$

3.2.3 ทำการเปรียบเทียบและสรุปผลในรูปตาราง

เมื่อได้ค่า Deviance ของแต่ละวิธีแล้ว นำผลที่ได้มาสรุปลงในตารางเพื่อแสดงการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์และเลือกตัวแบบให้เหมาะสมกับลักษณะข้อมูล

วิธีไคกำลังสองต่ำสุด

เริ่มต้น

ประมาณค่าพารามิเตอร์ $\hat{\beta}$ ด้วยวิธี MCS

$$b_{GLS} = [\mathbf{X'WX}]^{-1}\mathbf{X'W}g(\tilde{p}_i)$$

$$\text{logit } w_i = n_i\tilde{p}_i(1-\tilde{p}_i)$$
$$\text{probit } w_i = n_i\phi(\hat{z}_i)^2/\tilde{p}_i(1-\tilde{p}_i)$$
$$\text{clog - log } w_i = \{(\partial\hat{\eta}_i/\partial\hat{\mu}_i)^2 n_i\tilde{p}_i(1-\tilde{p}_i)\}^{-1},$$
$$\partial\hat{\eta}_i/\partial\hat{\mu}_i = \{-\log(1-\tilde{p}_i)n_i(1-\tilde{p}_i)\}^{-1}$$

คำนวณค่าความน่าจะเป็นของตัวแบบ โลจิท	คำนวณค่าความน่าจะเป็นของตัวแบบโพรบิท	คำนวณค่าความน่าจะเป็นของตัวแบบคอมพลีเมนทารีล็อก-ล็อก
$p(x_i) = \dfrac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$	$p(x_i) = \Phi(\beta_0+\beta_1 x_i)$	$p(x_i) = 1-\exp[-\exp(\beta_0+\beta_1 x_i)]$

คำนวณค่า ประมาณ(ค่าพยากรณ์) ของตัวแบบ

$$\hat{y} = np$$

คำนวณค่า Deviance

พิมพ์ผลลัพธ์

หยุด

บทที่ 4
ผลการวิเคราะห์ข้อมูล

งานวิจัยครั้งนี้เป็นการศึกษาวิธีการประมาณค่าพารามิเตอร์ของ ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก เพื่อเปรียบเทียบความถูกต้องของการประมาณค่าพารามิเตอร์ ซึ่งวิธีการประมาณพารามิเตอร์ของตัวแบบทั้ง 3 ตัวแบบมีวิธี 3 วิธีดังนี้

1. วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก(WLS)
2. วิธีภาวะน่าจะเป็นสูงสุด (MLE)
3. วิธีไคกำลังสองต่ำสุด (MCS)

โดยจะเปรียบเทียบเพื่อใช้ในการตัดสินว่าวิธีการใดเป็นวิธีที่ประมาณค่าพารามิเตอร์ที่ดีที่สุด และ ตัวแบบใดเป็นตัวแบบที่เหมาะสมกับลักษณะของข้อมูลมากที่สุด โดยใช้เกณฑ์การตัดสินใจ คือ ค่า Deviance ประกอบการตัดสินใจ โดยที่วิธีการประมาณค่าพารามิเตอร์ใดภายใต้ตัวแบบใดให้ค่า Deviance น้อยที่สุดเป็นวิธีประมาณค่าพารามิเตอร์ภายใต้ตัวแบบที่เหมาะสมกับลักษณะข้อมูลมากที่สุด

เพื่อความสะดวกในการอธิบาย จะใช้สัญลักษณ์แทนความหมายต่างๆ ดังนี้

Logit แทน ตัวแบบโลจิท
Probit แทน ตัวแบบโพรบิท
Clog-log แทน ตัวแบบคอมพลีเมนทารีล็อก-ล็อก
WLS แทน วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก
MLE แทน วิธีภาวะน่าจะเป็นสูงสุด
MCS แทน วิธีไคกำลังสองต่ำสุด

ตารางที่ 4.1 จำนวนผู้ได้รับทดสอบเซรุ่มที่ให้ผลบวก และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

กลุ่มอายุ	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
				Logit			Probit			C log- log		
				WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
0-11 เดือน	0.5	10	3	1.4	3.1	3.4	1.4	3.2	3.3	1.4	3.5	3.9
1-2 y	1.5	10	1	1.6	3.3	3.5	1.6	3.3	3.5	1.5	3.6	4
2-4 y	3	29	5	5.3	10.5	11.1	5.4	10.5	10.9	5.1	11.1	12.4
5-9 y	7	69	39	18.4	30.7	31.9	19	30.6	31.3	17	30.8	33.5
10-14 y	12	51	31	20.5	28.2	28.8	20.9	27.9	28.2	18.7	27.3	28.9
15-19 y	17	15	8	8.3	9.8	9.9	8.3	9.7	9.8	7.6	9.4	9.7
≥ 20	30	108	91	93	92.4	91.8	93.2	92.3	92.1	99.9	92.1	91.6

ตารางที่ 4.2 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่1

วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	38.4522	3.06E-07	35.86464	1.01E-06	53.93725	2.16E-10
MLE	13.76141**	0.017198	13.95967	0.015868	15.43142	0.00867
MCS	14.01068	0.015542	14.02182	0.015471	16.49553	0.005563

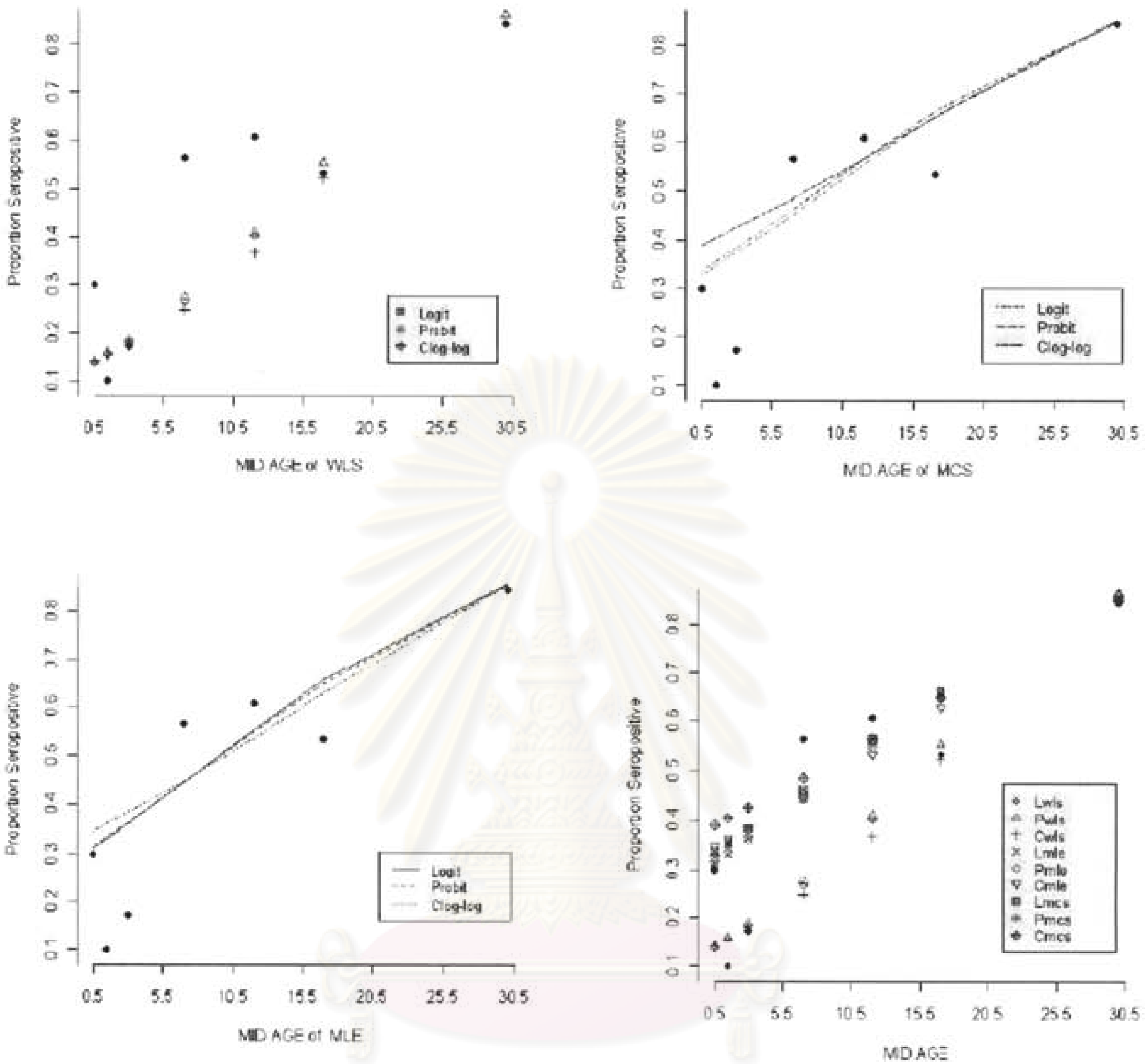
องศาความเป็นอิสระ df = 5

ข้อมูลผู้อาศัยในหมู่บ้าน Amozonas ประเทศบราซิล ปี 1971 ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ เพื่อตรวจสอบในช่วงเวลาในการฉีดเซรุ่มป้องกันมาลาเรีย ว่าเซรุ่มที่ให้ผลบวกหรือไม่ให้ผลบวก ข้อมูลจาก Draper,Voller and Carpenter(1972) ซึ่ง Draper ได้ใช้ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ในการวิเคราะห์ข้อมูล โดยมีค่ากลางของกลุ่มอายุ (X)เป็นตัวประกอบหนึ่งที่เป็นตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ ผู้ได้รับเซรุ่มที่ให้ผลบวกและไม่ได้ให้ผลบวกเมื่อค่ากลางของอายุต่างกัน ข้อมูลคือ ความถี่ของผู้ที่ได้รับเซรุ่มที่ให้ผลบวกและไม่ได้ให้ผลบวก ณ ค่ากลางของกลุ่มอายุที่ระดับต่างกัน พบว่าค่าประมาณที่ได้จากตัวแบบ โลจิท ตัวแบบโพรบิท และ ตัว

แบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และวิธีไคกำลังสองต่ำสุด (MCS)จำแนกตามค่ากลางของกลุ่มอายุ แสดงไว้ในตารางที่ 4.1

การทดสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ดังตารางที่ 4.2 โดยค่าสถิติ Deviance ของตัวแบบโลจิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 13.76141 น้อยกว่าทุกกรณี แต่ค่า Deviance เทียบกับค่า χ^2_5 ได้ค่า p-value ที่ระดับองศาอิสระเท่ากับ 5 และความน่าจะเป็นในการยอมรับตัวแบบโลจิทของข้อมูลชุดที่ 1 นี้มีค่าเท่ากับ 0.017198 น้อยกว่าระดับนัยสำคัญ 0.05 จึงถือได้ว่าตัวแบบโลจิทยังไม่มีความเหมาะสมกับข้อมูลชุดที่ 1 นี้

จากรูปที่ 4.1 แสดงการพล็อตความน่าจะเป็นผู้ป่วยได้รับเซรุ่มที่ให้ผลบวกตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่ออายุเพิ่มขึ้นทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.1 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตไม่ได้เป็นกราฟรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ห่างจากจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่า deviane มีค่ามาก กล่าวได้ว่าตัวแบบโลจิทที่ประมาณได้ยังไม่มีความเหมาะสมพอกับข้อมูลชุดที่ 1



รูปที่ 4.1 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลที่ 1

ตารางที่ 4.3 ข้อมูลผู้สูบบุหรี่ที่มีอาการหอบ และ ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

กลุ่มอายุ	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
				Logit			Probit			C log- log		
				WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
20-24y	22	1952	104	105.9	112.9	113.1	101.7	102.3	102.4	108	120.6	121.2
25-29y	27	1791	128	133.1	140.3	140.6	134.3	135	135.1	132.9	145.7	146.3
30-34y	32	2113	231	213.5	222.6	222.9	221.6	222.5	222.6	209.7	225.5	226.3
35-39y	37	2783	378	378.3	390.3	390.7	396.6	398	398.1	367.6	387.9	389.1
40-44y	42	2274	442	410.4	419.1	419.5	428.3	429.6	429.7	397.2	411.6	412.7
45-49y	47	2393	593	563.8	570.4	570.7	579.9	581.3	581.4	548.2	558.3	559.4
50-54y	52	2090	649	629.9	632	632.3	634.8	636.1	636.2	621.1	622.2	623.1
55-59y	57	1750	631	658.8	656.6	656.7	649.7	650.8	650.8	664.9	656.4	657
60-64y	62	1136	504	520.3	515.9	515.8	503.3	503.9	503.9	541.6	528.2	528.5

องศาความเป็นอิสระ df = 7

ตารางที่ 4.4 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 2

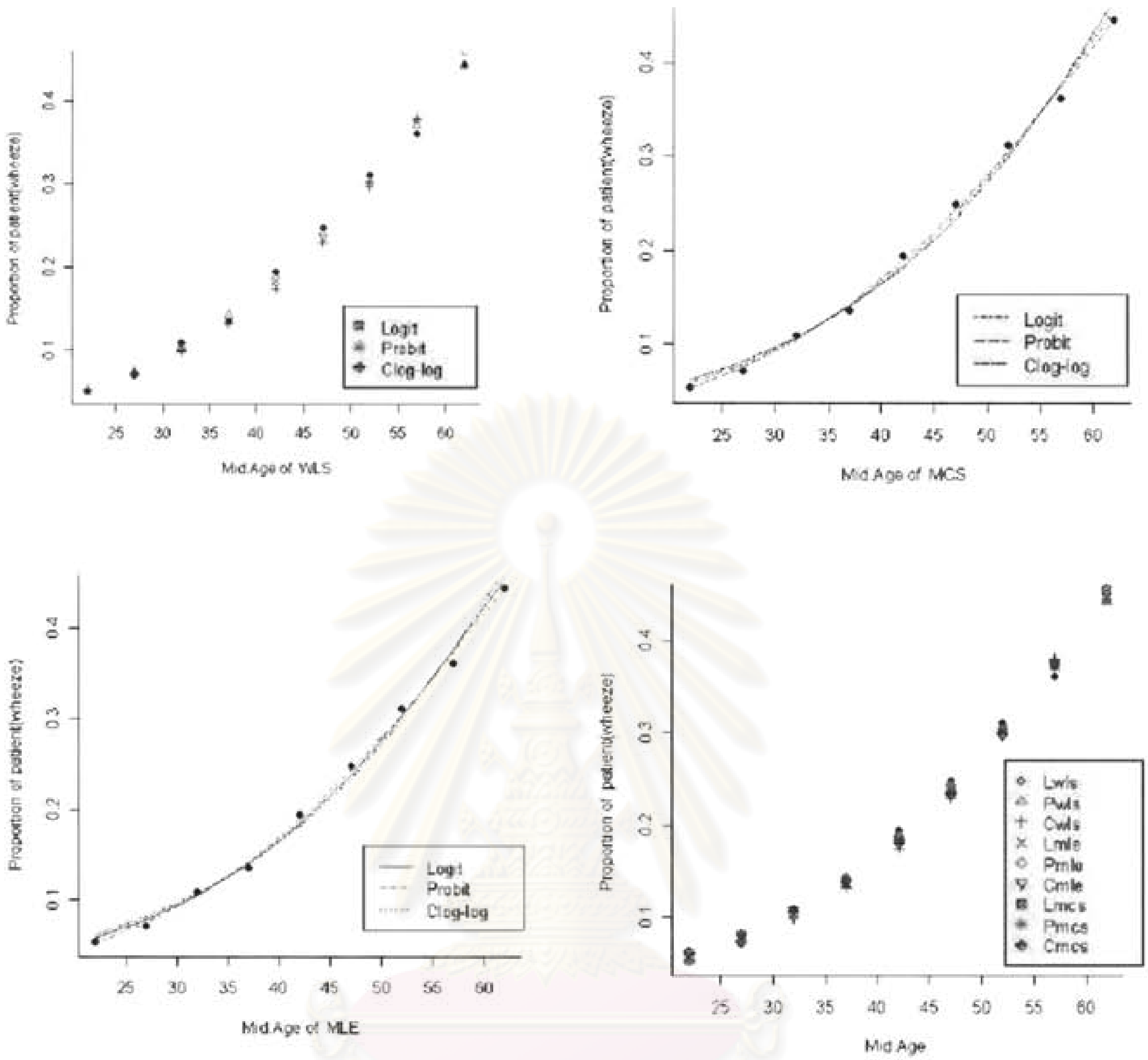
วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	10.36431	1.69E-01	4.09082	7.69E-01	23.20187	1.57E-03
MLE	8.19993	0.315230	4.05551**	0.77336	16.15742	0.02372
MCS	8.20316	0.31502	4.05565*	0.77335	16.1799	0.02352

ข้อมูลผู้สูบบุหรี่ที่ปราศจากสารกัมมันตภาพรังสีที่มีอายุระหว่าง 20-64 ปี ของบริษัท Coalminers ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ ว่ามีอาการหอบหรือไม่มีอาการหอบ เมื่อทำงานใน Coalminers ข้อมูลชุดนี้ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล ข้อมูลจาก Ashford and Sowden(1970) ได้ซึ่งค่ากลางของกลุ่มอายุ (X)เป็นตัวประกอบหนึ่งที่เป็นตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ ผู้สูบบุหรี่มีอาการหอบ และ ไม่มีอาการหอบ เมื่อค่ากลางของอายุต่างกัน ข้อมูล คือ ความถี่ของผู้สูบบุหรี่มีอาการหอบและไม่มีอาการหอบ ณ ค่ากลางของกลุ่มอายุที่

ระดับต่างกัน พบว่าค่าประมาณที่ได้จาก ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุดMLE และวิธีไคกำลังสองต่ำสุด (MCS)จำแนกตามค่ากลางของกลุ่มอายุ แสดงไว้ในตารางที่ 4.3

การทดสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ดังตารางที่ 4.4 โดยค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 4.05551 ให้ค่า P-value เท่ากับ 0.77336 ที่องศาความเป็นอิสระเท่ากับ 7 ความน่าจะเป็นในการยอมรับตัวแบบโพรบิทมีมากกว่าความน่าจะเป็นตัวแบบโลจิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก รองลงมา คือ ตัวแบบโพรบิทภายใต้วิธี MCS ให้ค่า Deviance เท่ากับ 4.05565 ให้ค่า P-value เท่ากับ 0.77335 ซึ่งค่าประมาณของตัวแบบโพรบิท ภายใต้วิธี MLE และวิธี MCS ให้ค่าประมาณที่ใกล้เคียงกันมาก เป็นไปตามที่ Berkson(1955) กล่าวไว้ว่าค่าประมาณพารามิเตอร์จาก MCS ให้ค่าประมาณได้ดีเท่ากับวิธี MLE จากข้อมูลผู้ที่สูบบุหรี่ที่มีอาการหอบชุดนี้ ควรเลือกตัวแบบโพรบิทภายใต้วิธี MLE เพราะให้ค่าสถิติ Deviance น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ จึงถือได้ว่าตัวแบบโพรบิทภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดนี้มากที่สุด และให้ค่าประมาณที่ได้จากตัวแบบใกล้เคียงกับค่าสังเกตจากข้อมูลจริงมากที่สุด ซึ่งก่อนหน้านี้ Ashford and Sowden(1970) ได้ทำการวิเคราะห์ด้วยตัวแบบโลจิท จะเห็นได้ว่าค่า Deviance จากวิธี MLE ของข้อมูลชุดนี้ มีค่าเท่ากับ 8.19993 ให้ค่า P-value เท่ากับ 0.3152 หรือกล่าวได้ว่าความน่าจะเป็นในการยอมรับตัวแบบมีแค่ 31.52 เปอร์เซ็นต์ น้อยกว่าตัวแบบโพรบิทเมื่อใช้วิธี MLE ดังนั้นจึงควรเลือกใช้ตัวแบบโพรบิท ภายใต้วิธี MLE เพราะมีความเหมาะสมกับข้อมูลชุดที่ 2

จากรูปที่ 4.2 แสดงการพล็อตความน่าจะเป็นผู้ป่วยที่สูบบุหรี่ที่มีอาการหอบ ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่ออายุเพิ่มขึ้นทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.2 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นกราฟรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่า deviane มีค่าน้อย กล่าวได้ว่าตัวแบบโพรบิทที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 2



รูปที่ 4.2 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 2

ตารางที่ 4.5 ข้อมูลของเพศชายที่เป็นโรคหัวใจ และ ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

ความดันโลหิต	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
				Logit			Probit			C log- log		
				WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
<117	111.5	156	3	3.6	5.2	5.6	3.5	5	5.2	3.6	5.3	5.7
117-126	121.5	252	17	7.7	10.6	11.3	7.9	10.5	10.9	7.7	10.7	11.4
127-136	131.5	284	12	11.7	15.1	15.9	12.2	15.2	15.6	11.6	15.1	15.9
137-146	141.5	271	16	14.9	18.1	18.8	15.8	18.4	18.8	14.7	18	18.8
147-156	151.5	139	12	10.2	11.6	11.9	10.7	11.8	12	10	11.5	11.9
157-166	161.5	85	8	8.2	8.9	9	8.5	9	9	8.1	8.8	8.9
167-186	176.5	99	16	14.3	14.2	14.1	14.2	14.1	14	14.3	14.2	14.1
>186	191.5	43	8	9	8.4	8.2	8.6	8	7.9	9.2	8.5	8.3

องศาความเป็นอิสระ df = 6

ตารางที่ 4.6 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 3

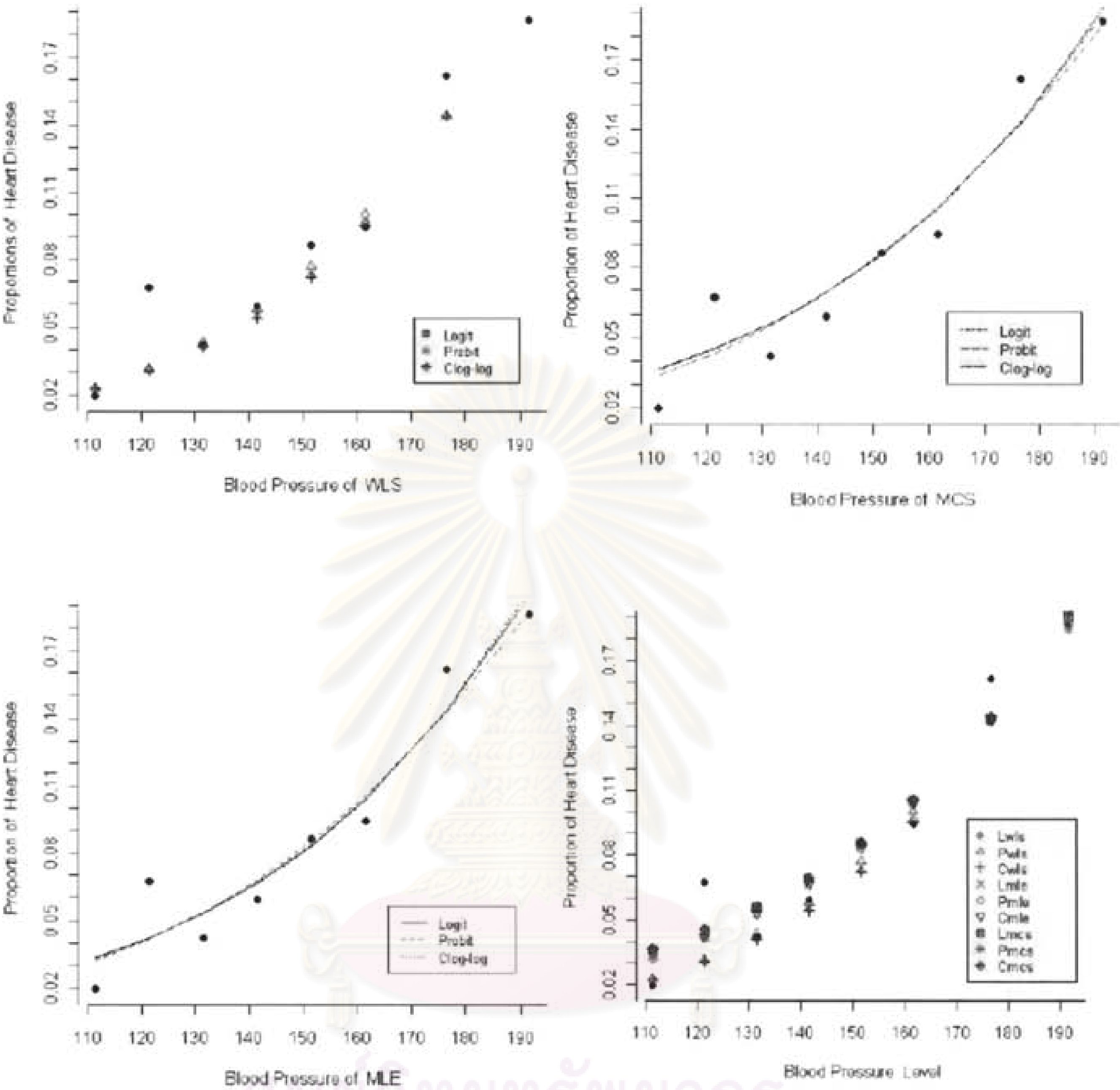
วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	9.53954	0.14543	8.73983	0.18875	9.74804	0.13567
MLE	5.90916*	0.43344	6.05616	0.41693	5.88427**	0.43628
MCS	6.08157	0.41411	6.10693	0.41132	6.08087	0.41419

ข้อมูลผู้อาศัยเพศชายอายุ 40-59 ปี ในเมือง 2 เมือง ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความดันโลหิต เพื่อตรวจสอบในช่วง 6 ปี ต่อเนื่องกันว่าเป็นโรคหัวใจ หรือไม่เป็นโรคหัวใจ โดยได้ทำการวิเคราะห์ข้อมูลชุดนี้ด้วยตัวแบบโลจิท ข้อมูลจาก Cornfield (1962) และ Agresti (1990) ซึ่งค่ากลางของความดันโลหิต (X)เป็นตัวประกอบหนึ่งที่เป็นตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ ผู้ที่เป็นโรคหัวใจหรือ ไม่เป็นโรคหัวใจ เมื่อค่ากลางของความดันโลหิตต่างกัน ข้อมูลคือ ความถี่ของผู้ที่เป็นโรคหัวใจหรือ ไม่เป็นโรคหัวใจ ณ ค่ากลางของความดันโลหิตที่ระดับต่างกัน พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-

ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุดMLE และวิธีไคกำลังสองต่ำสุด (MCS)จำแนกตามค่ากลางของกลุ่มอายุ แสดงไว้ในตารางที่ 4.5

การทดสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ดังตารางที่ 4.6 โดยค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 5.88427 ให้ค่า P-value เท่ากับ 0.43628 ที่ค่าองศาความเป็นอิสระเท่ากับ 6 รองลงมาคือ ค่าสถิติ Deviance ของตัวแบบโลจิทภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 5.9092 ให้ค่า P-value เท่ากับ 0.4334 จะเห็นว่าความน่าจะเป็นในการยอมรับตัวแบบคอมพลีเมนทารีล็อก-ล็อก มีมากกว่า ความน่าจะเป็นในการยอมรับตัวแบบโลจิท และ ตัวแบบโพรบิท จากข้อมูลชุดนี้ควรเลือกตัวแบบ คอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE เพราะให้ค่าสถิติ Deviance น้อยที่สุดจากทุกกรณีของ วิธีการประมาณค่าพารามิเตอร์ จึงถือได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE และ ความน่าจะเป็นในการยอมรับตัวแบบ มีถึง 43.63% กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก มี ความเหมาะสมกับข้อมูลชุด 3 นี้มากที่สุด ภายใต้ระดับนัยสำคัญ 0.05

จากรูปที่ 4.3 แสดงการพล็อตความน่าจะเป็นผู้ป่วยที่เป็นโรคหัวใจ ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อความดันโลหิตเพิ่มขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.3 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นปลายหางของรูปตัวเอส เนื่องจากว่าความน่าจะเป็นของการพล็อตข้อมูลชุด 3 นี้มีค่าสูงสุดที่ 0.2 ทำให้มองกราฟออกมาไม่เป็นรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่า Deviane มีค่าน้อย กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 3



รูปที่ 4.3 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 3

ตารางที่4.7 จำนวนการตายของแมลงเมื่อได้รับสารพิษแต่ละระดับ และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

log-dose	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
			Logit			Probit			C log- log		
			WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
0.41	50	6	6.4	6.3	6.3	6.4	6.3	6.3	6.9	8.4	8.6
0.58	48	16	15.5	15.5	15.5	15.8	15.8	15.8	13.7	15.5	15.7
0.71	46	24	24.9	25.1	24.9	24.7	24.8	24.8	21.5	23	23.2
0.89	49	42	39.4	39.7	39.5	39.1	39.3	39.2	38.2	38.4	38.5
1.01	50	44	45.2	45.4	45.3	45.4	45.6	45.5	46.6	46.3	46.3

องศาความเป็นอิสระ df = 3

ตารางที่ 4.8 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 4

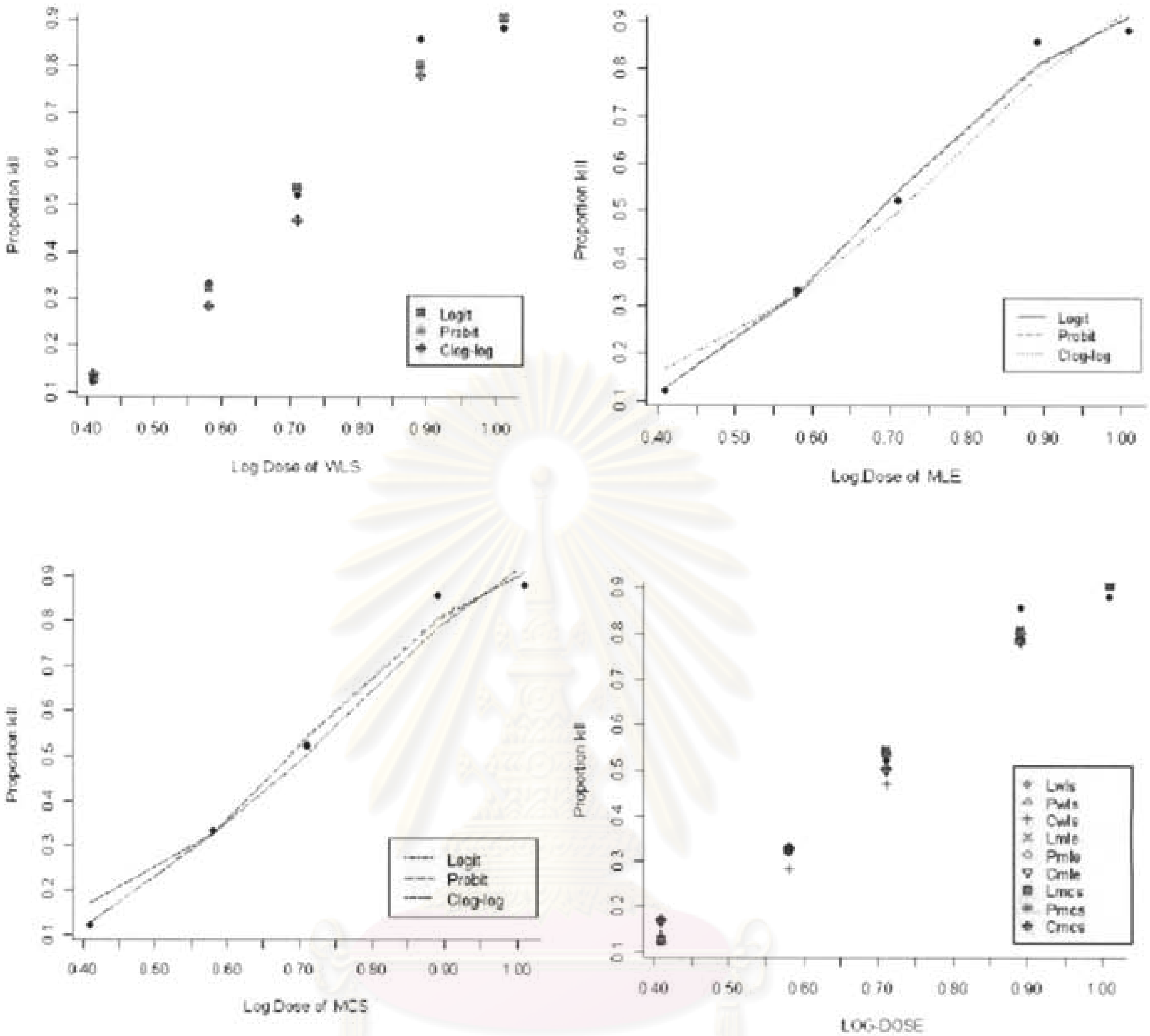
วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	1.37558	7.11E-01	1.65452	6.47E-01	4.91120	1.78E-01
MLE	1.34806**	0.71775	1.63864	0.65066	4.03547	0.25766
MCS	1.36100*	0.71470	1.64286	0.64971	4.05094	0.25602

ข้อมูลปริมาณความเข้มข้นของสารพิษที่มีผลต่อการตายของแมลง ข้อมูลชุดนี้ได้ทำการวิเคราะห์ด้วยตัวแบบโพรบิท ข้อมูลจาก Martin(1942) และ Finney(1971) ปริมาณความเข้มข้นของสารพิษ log-dose ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ การตายและไม่ตายของแมลงเมื่อได้รับความเข้มข้นของสารพิษในปริมาณต่าง ๆ ข้อมูล คือ จำนวนความถี่ของการตายและไม่ตายของแมลง ณ ระดับความเข้มข้นของสารพิษในความเข้มข้นต่าง ๆ พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุดMLE และวิธีไคกำลังสองต่ำสุด (MCS) จำแนกตามระดับของ log-dose แสดงไว้ในตารางที่ 4.7

การทดสอบความเหมาะสมของตัวแบบ ภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.8 โดยค่าสถิติ Deviance ของตัวแบบโลจิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 1.34806 ให้ค่า P-value เท่ากับ 0.71775 ที่องศาความเป็นอิสระเท่ากับ 3 ความน่าจะเป็นในการยอมรับตัวแบบโลจิท มีมากกว่าความน่าจะเป็น ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก - ล็อก รองลงมาคือ ตัวแบบโลจิทภายใต้วิธี MCS ให้ค่า Deviance เท่ากับ 1.36100 ซึ่งค่าประมาณของตัวแบบโลจิทภายใต้วิธี MLE และวิธี MCS ให้ค่าประมาณที่ใกล้เคียงกันมาก เป็นไปตามที่ Berkson(1955) กล่าวไว้ว่าค่าประมาณพารามิเตอร์จาก MCS ให้ค่าประมาณได้ดีเท่ากับวิธี MLE จากข้อมูลชุดนี้ ควรเลือกตัวแบบโลจิทภายใต้วิธีการประมาณค่า พารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด MLE เพราะให้ค่าสถิติ Deviance น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ จึงถือได้ว่าตัวแบบโลจิทภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดนี้มากที่สุด และให้ค่าประมาณที่ได้จากตัวแบบใกล้เคียงกับค่าสังเกตจากข้อมูลจริงมากที่สุด ภายใต้ระดับนัยสำคัญ 0.05

ข้อมูลชุดที่ 4 นี้หลังจากทำการปรับค่าตัวแปรอธิบายแล้ว ผลที่ได้คือ ตัวแบบโลจิท มีความเหมาะสมมากกว่าตัวแบบโพรบิทที่ยังไม่ได้ทำการปรับข้อมูลตัวแปรอธิบาย เหตุผลที่เป็นเช่นนี้เพราะว่า ขนาดความแปรปรวนขึ้นอยู่กับขนาดของข้อมูล และความแตกต่างของค่าพารามิเตอร์ที่ประมาณได้จะขึ้นอยู่กับขนาดของ σ เท่านั้น ดังนั้นจะเห็นได้ว่าตัวแปรอธิบายมีผลต่อการเลือกลิงค์ฟังก์ชัน

จากรูปที่ 4.4 แสดงการพล็อตความน่าจะเป็นการตายของแมลง ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อปริมาณความเข้มข้นของสารพิษเพิ่มขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.4 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Deviane มีค่าน้อย กล่าวได้ว่าตัวแบบโลจิท ที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 4



รูปที่ 4.4 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 4

ตารางที่ 4.9 จำนวนการตายของแมลง และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธีของข้อมูลชุดที่ 5

CONC	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
			Logit			Probit			C log- log		
			WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
0.0018	10	1	1.1	1.4	1.4	1.1	1.3	1.3	1.2	1.7	1.8
0.0022	10	3	2.4	2.7	2.8	2.5	2.8	2.8	2.2	2.8	2.9
0.0026	10	5	4.5	4.7	4.7	4.5	4.7	4.7	4	4.4	4.5
0.003	10	7	6.6	6.8	6.8	6.6	6.7	6.7	6.4	6.5	6.5
0.0034	10	8	8.3	8.3	8.3	8.3	8.3	8.3	8.7	8.5	8.5

องศาความเป็นอิสระ df = 3

ตารางที่ 4.10 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 5

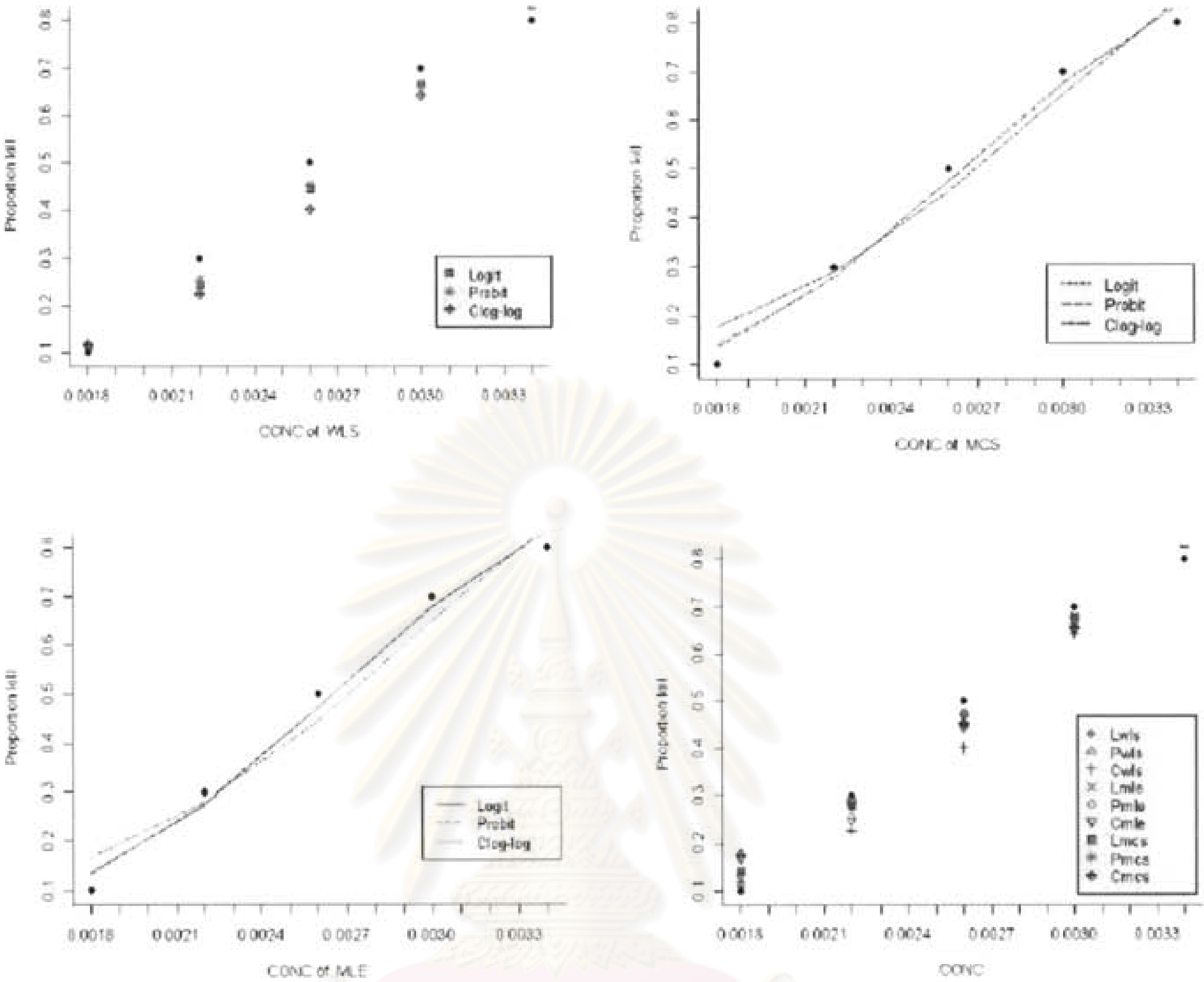
วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	0.43638	0.93263	0.35387	0.94960	1.32623	0.72292
MLE	0.29161	0.96160	0.27191**	0.96522	0.79039	0.85176
MCS	0.29382	0.96119	0.27224*	0.96516	0.80553	0.84814

ข้อมูลปริมาณความเข้มข้นของสารพิษที่มีผลต่อการตายของแมลง ข้อมูลชุดนี้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิสติคในการวิเคราะห์ และทำการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนักว่ามีความเหมาะสมกับข้อมูลทางชีววิทยา ข้อมูลจาก Muhammad(1990) ปริมาณความเข้มข้นของสารพิษ (X) ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ การตายและไม่ตายของแมลงเมื่อได้รับความเข้มข้นของสารพิษในปริมาณต่าง ๆ ข้อมูล คือ จำนวนความถี่ของการตายและไม่ตายของแมลง ณ ระดับความเข้มข้นของสารพิษในความเข้มข้นต่าง ๆ พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วง

น้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และวิธีไคกำลังสองต่ำสุด (MCS) จำแนกตามระดับของความเข้มข้นของสาร แสดงไว้ในตารางที่ 4.9

การทดสอบความเหมาะสมของตัวแบบ ภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.10 โดยค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 0.27191 ให้ค่า P-value เท่ากับ 0.96522 ความน่าจะเป็นในการยอมรับตัวแบบโพรบิท 96.522% มีมากกว่าความน่าจะเป็นตัวแบบโลจิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก รองลงมา คือ ตัวแบบโพรบิท ภายใต้วิธี MCS ให้ค่า Deviance = 0.27224 ให้ค่า P-value เท่ากับ 0.96516 ความน่าจะเป็นในการยอมรับตัวแบบโพรบิทด้วยวิธี MCS มี 96.516% มีมากกว่าความน่าจะเป็นตัวแบบโลจิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ซึ่งค่าประมาณของตัวแบบโพรบิท ภายใต้วิธี MLE และวิธี MCS ให้ค่าประมาณที่ใกล้เคียงกันมาก เป็นไปตามที่ Berkson(1955) กล่าวไว้ว่าค่าประมาณพารามิเตอร์จาก MCS ให้ค่าประมาณได้ดีเท่ากับวิธี MLE จากข้อมูลชุดนี้ควรเลือกตัวแบบโพรบิทภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) เพราะให้ค่าสถิติ Deviance น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ และความน่าจะเป็นในการยอมรับตัวแบบมีมากถึง 96.52% จึงถือได้ว่าตัวแบบโพรบิทภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุด 5 นี้มากที่สุด และให้ค่าประมาณที่ได้จากตัวแบบใกล้เคียงกับค่าสังเกตจากข้อมูลจริงมากที่สุด ภายใต้ระดับนัยสำคัญ 0.05

จากรูปที่ 4.5 แสดงการพล็อตความน่าจะเป็นของการตายของแมลง ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อปริมาณความเข้มข้นของสารพิษเพิ่มขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.5 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Deviane มีค่าน้อย กล่าวได้ว่าตัวแบบโพรบิท ที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 5



รูปที่ 4.5 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณ
ค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 5

ตารางที่ 4.11 จำนวนการตายของแมลง และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธีของข้อมูลชุดที่ 6

log-dose	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
			Logit			Probit			C log- log		
			WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
0.72	58	3	3.9	2.8	3.5	4.2	2.7	3	4.4	5.6	6.6
0.8	61	19	14.9	10.7	11.8	15.7	11.7	12.1	10.9	12.9	14.4
0.87	63	16	34.4	26.7	27.3	32.9	27.5	27.6	22.1	24.9	26.3
0.93	59	37	46.4	40.2	39.7	44	39.8	39.4	33.8	36.3	37
0.98	57	49	51.5	47.8	47	50.1	47.4	46.9	44.3	45.8	45.8
1.02	55	54	52.4	50.3	49.6	51.8	50.3	49.9	49.8	50.5	50.1
1.07	57	55	55.9	54.9	54.4	55.9	55.3	55	56.1	56.2	56
1.1	61	60	60.3	59.6	59.3	60.5	60.1	59.9	60.8	60.8	60.7

องศาความเป็นอิสระ df = 6

ตารางที่ 4.12 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 6

วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	34.86793	0.00000	27.47521	0.00012	21.15839	0.00172
MLE	20.37758	0.00237	19.65566*	0.00319	19.12758**	0.00395
MCS	21.02282	0.00182	19.82097	0.00298	19.86979	0.00292

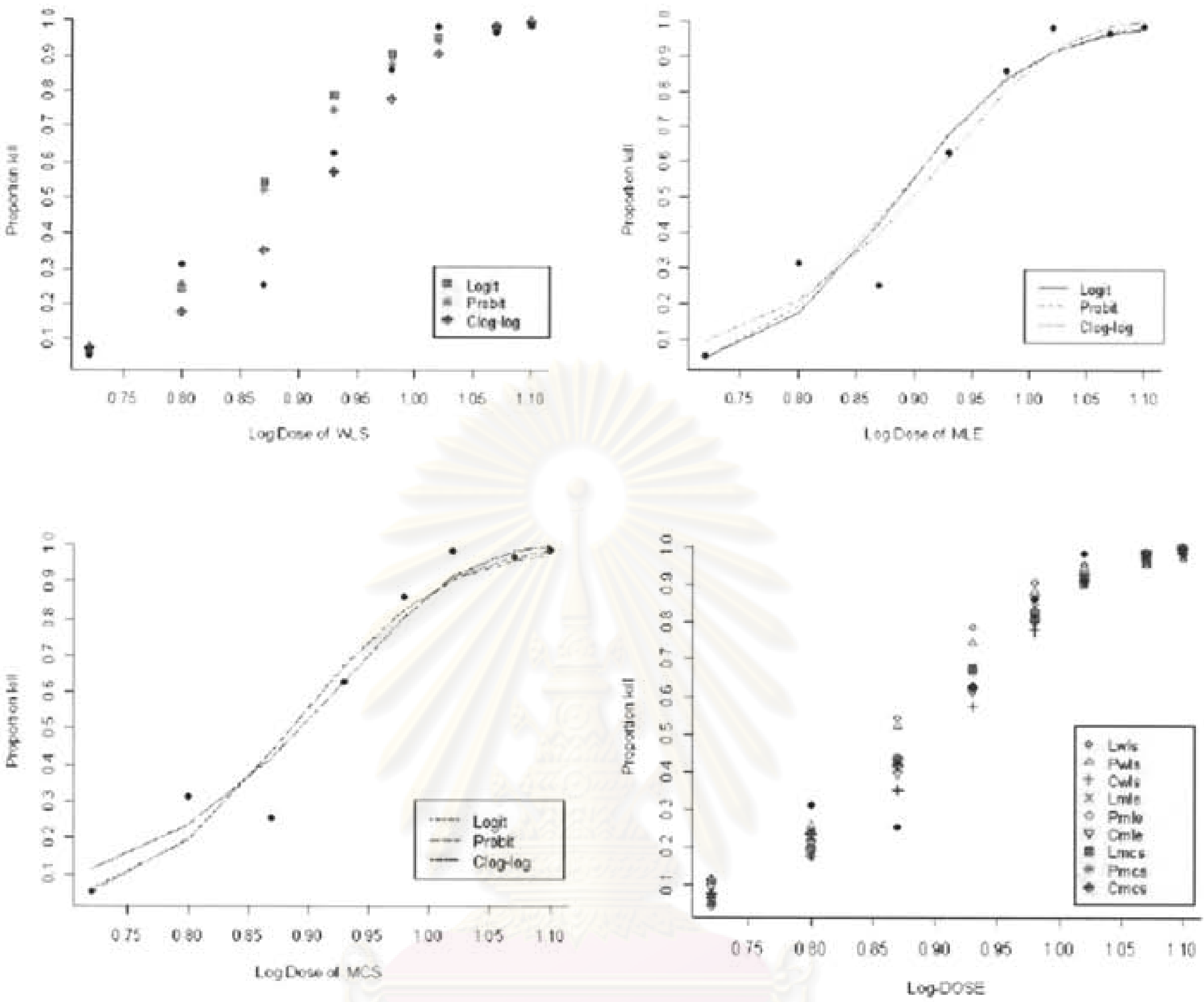
ข้อมูลความเข้มข้นของแอมโมเนีย ที่มีผลต่อการตายของแมลง ข้อมูลชุดนี้ใช้ตัวแบบโพรบิทในการวิเคราะห์ข้อมูล ข้อมูลจาก Strand(1930) จำแนกออกตามความเข้มข้นของสารพิษ (X) ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ การตายและไม่ตายของแมลงเมื่อได้รับความเข้มข้นของสารพิษในปริมาณต่าง ๆ ข้อมูล คือ จำนวนความถี่ของการตายและไม่ตายของแมลง ณ ระดับความเข้มข้นของสารพิษในความเข้มข้นต่าง ๆ พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณ

ค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด (MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) จำแนกตามระดับของ log-dose แสดงไว้ในตารางที่ 4.11

การทดสอบความเหมาะสมของตัวแบบ ภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.12 โดยค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 19.12758 ให้ค่า P-value เท่ากับ 0.00395 น้อยกว่าระดับนัยสำคัญ 0.05 รองลงมา คือ ตัวแบบโพรบิทภายใต้วิธี MLE ให้ค่า Deviance = 19.65566 นั่นหมายความว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ไม่มีความเหมาะสมกับข้อมูลชุดนี้ ถึงแม้ว่าจะให้ค่า Deviance น้อยกว่าทุกกรณีก็ตาม

ข้อมูลชุดที่ 6 นี้หลังจากทำการปรับค่าตัวแปรอธิบายแล้ว ผลที่ได้คือ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก มีความเหมาะสมมากกว่าตัวแบบโพรบิทที่ยังไม่ได้ทำการปรับข้อมูลตัวแปรอธิบาย เหตุผลที่เป็นเช่นนี้เพราะว่า ขนาดความแปรปรวนขึ้นอยู่กับขนาดของข้อมูล และ ความแตกต่างของค่าพารามิเตอร์ที่ประมาณได้จะขึ้นอยู่กับขนาดของ σ เท่านั้น ดังนั้นจะเห็นได้ว่าตัวแปรอธิบายมีผลต่อการเลือกลิงค์ฟังก์ชัน

จากรูปที่ 4.6 แสดงการพล็อตความน่าจะเป็นของการตายของแมลง ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อปริมาณความเข้มข้นของสารพิษเพิ่มขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.6 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ มีบางจุดที่ห่างจากจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Devianc มีค่ามาก กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ที่ประมาณได้ด้วยวิธี MLE ยังไม่ความเหมาะสมกับข้อมูลชุดที่ 6



รูปที่ 4.6 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 6

ตารางที่ 4.13 จำนวนหมุดที่เกิดความเสียหาย และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

Pressure Load	log-load	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
				Logit			Probit			C log- log		
				WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
2500	7.824046	50	10	9	8.4	8.4	8.8	8.2	8.2	9.6	9.3	9.3
2700	7.901007	70	17	17	16.1	16.2	17.1	16.2	16.2	17.3	16.8	16.8
2900	7.972466	100	30	31.5	30.3	30.3	31.8	30.6	30.6	30.8	30.2	30.2
3100	8.039157	60	21	23.5	22.8	22.8	23.6	23	23	22.5	22.2	22.2
3300	8.101678	40	18	18.7	18.3	18.3	18.7	18.4	18.4	17.8	17.7	17.7
3500	8.160518	85	43	46	45.4	45.4	45.9	45.5	45.5	44.1	44	44
3700	8.216088	90	54	54.8	54.4	54.4	54.6	54.4	54.4	53.2	53.3	53.4
3900	8.268732	50	33	33.5	33.4	33.4	33.3	33.3	33.3	33.1	33.3	33.3
4100	8.318742	80	60	57.8	57.8	57.8	57.6	57.8	57.7	58.2	58.7	58.7
4300	8.36637	65	51	49.9	50	50	49.9	50.1	50	51.2	51.6	51.6

องศาความเป็นอิสระ df = 8

ตารางที่ 4.14 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 7

วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	1.61067	0.99907	1.74884	0.98777	0.49184	0.99987
MLE	1.36195	0.99477	1.51713	0.99242	0.40519**	0.99994
MCS	1.36298	0.99476	1.51736	0.99242	0.40530*	0.99994

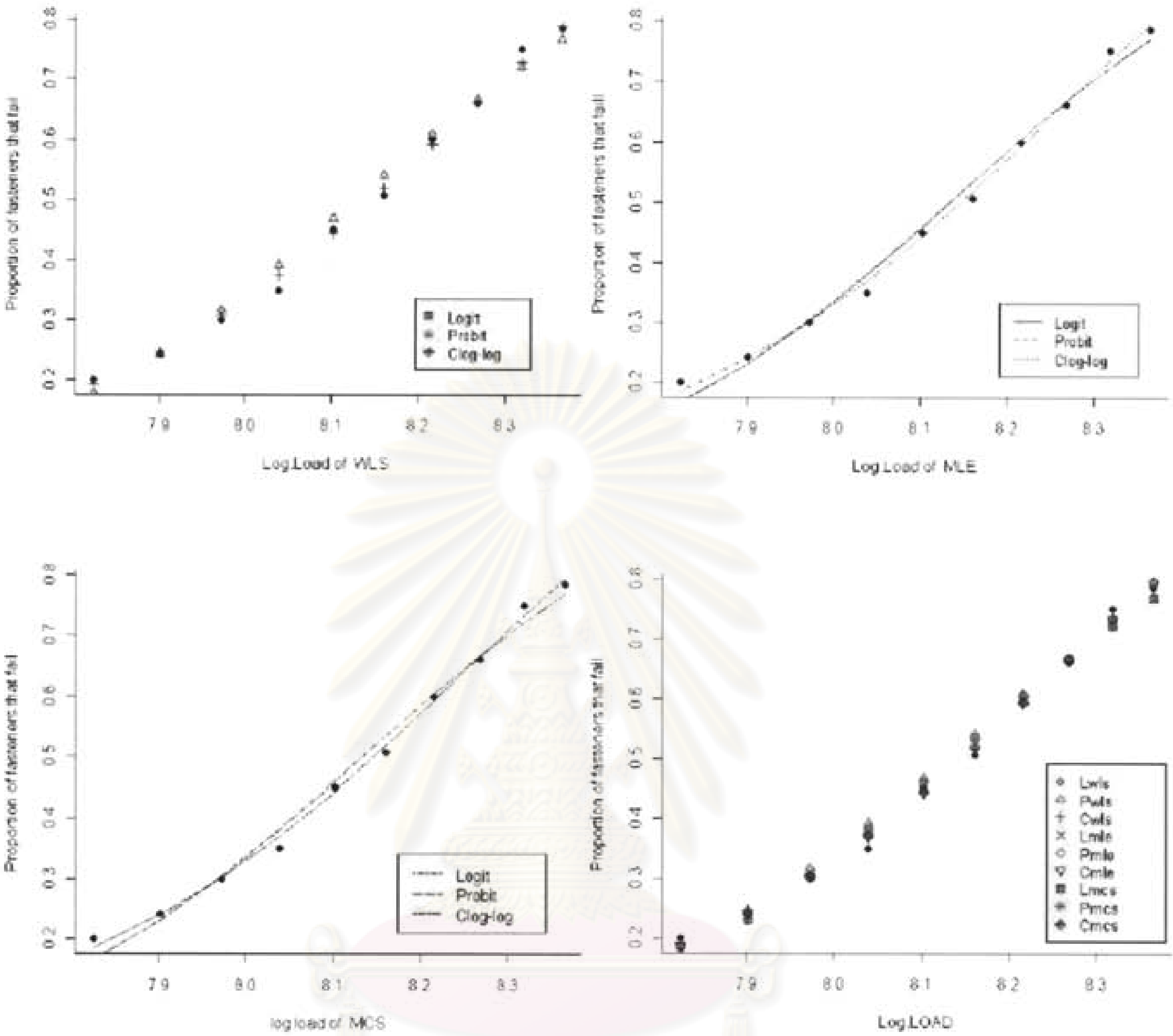
ข้อมูลการเสียหายของหมุดเจาะบนเครื่องบิน เมื่อระดับความกดอากาศเพิ่มขึ้นที่ละ 200 psi จาก 2500-4300 psi ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความกดอากาศ ว่าความกดอากาศที่ระดับต่างจะมีผลต่อการเสียหายหรือไม่เสียหายของหมุดเจาะบนเครื่องบิน ข้อมุลชุดนี้ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ ข้อมูลจาก Montgomery and Peck(1982) ทางผู้วิจัยได้ทำการเทค log เพื่อปรับตัวแปรอธิบายให้มีการแจกแจงแบบปกติ ระดับความกดอากาศ ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ หมุดเจาะเสียหายและไม่เสียหายเมื่อระดับความกดอากาศต่างระดับกัน

ข้อมูล คือ จำนวนความถี่ของหมุดเจาะที่เสียหายและไม่เสียหาย ณ ระดับความกดอากาศที่ระดับต่าง ๆ พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) จำแนกตามระดับความกดอากาศ แสดงไว้ในตารางที่ 4.13

การทดสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.14 โดยค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 0.40519 ให้ค่า P-value เท่ากับ 0.99994 ที่องศาความเป็นอิสระเท่ากับ 8 ความน่าจะเป็นในการยอมรับตัวแบบคอมพลีเมนทารีล็อก-ล็อก มีมากกว่าความน่าจะเป็นตัวแบบโลจิท และ ตัวแบบโพรบิท รองลงมาคือ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MCS ให้ค่า Deviance เท่ากับ 0.40530 ให้ค่า P-value เท่ากับ 0.99994 ซึ่งค่าประมาณของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MLE และวิธี MCS ให้ค่าประมาณและความน่าจะเป็นในการยอมรับตัวแบบที่ใกล้เคียงกันมาก จากข้อมูลทางวิศวกรรมชุดนี้ ควรเลือกตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MLE เพราะให้ค่าสถิติ Deviance น้อยที่สุด จากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ จึงถือได้ว่าตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดนี้มากที่สุด และให้ค่าประมาณที่ได้จากตัวแบบใกล้เคียงกับค่าสังเกตจากข้อมูลจริงมากที่สุด ภายใต้ระดับนัยสำคัญ 0.05

ข้อมูลชุดที่ 7 นี้หลังจากทำการปรับค่าตัวแปรอธิบายแล้ว ผลที่ได้คือ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก มีความเหมาะสมมากว่าตัวแบบโลจิทที่ยังไม่ได้ทำการปรับข้อมูลตัวแปรอธิบาย เหตุผลที่เป็นเช่นนี้เพราะว่า ขนาดความแปรปรวนขึ้นอยู่กับขนาดของข้อมูล และความแตกต่างของค่าพารามิเตอร์ที่ประมาณได้จะขึ้นอยู่กับขนาดของ σ เท่านั้น ดังนั้นจะเห็นได้ว่าตัวแปรอธิบายมีผลต่อการเลือกลิงค์ฟังก์ชัน

จากรูปที่ 4.7 แสดงการพล็อตความน่าจะเป็นของหมุดเจาะที่เกิดความเสียหาย ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อความกดอากาศเพิ่มขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.7 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นรูปตัวเอส การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Deviane มีค่าน้อย กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 7



รูปที่ 4.7 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 7

ตารางที่ 4.15 ข้อมูลการเลือก Reagan เป็นประธานาธิบดี และค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

Political views	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
			Logit			Probit			C log- log		
			WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
1	13	1	1.1	2.6	2.8	1.1	2.6	2.6	1.1	3.2	3.5
2	70	13	9.6	20.2	21.0	10.2	20.1	20.5	8.9	22.3	23.9
3	115	44	25.3	45.4	46.4	27.2	45.5	45.9	22.7	46.7	48.9
4	261	155	87.6	133.9	134.7	92.1	133.5	133.8	77.8	132.1	135.5
5	153	92	72.6	96.3	95.9	74.2	95.9	95.7	66.5	94.1	94.9
6	141	100	87.1	103.3	102.4	87.2	103.1	102.7	84.7	102.3	101.9
7	26	18	19.3	21.2	21.0	19.2	21.3	21.2	20.1	21.5	21.2

องศาความเป็นอิสระ df = 5

ตารางที่ 4.16 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 8

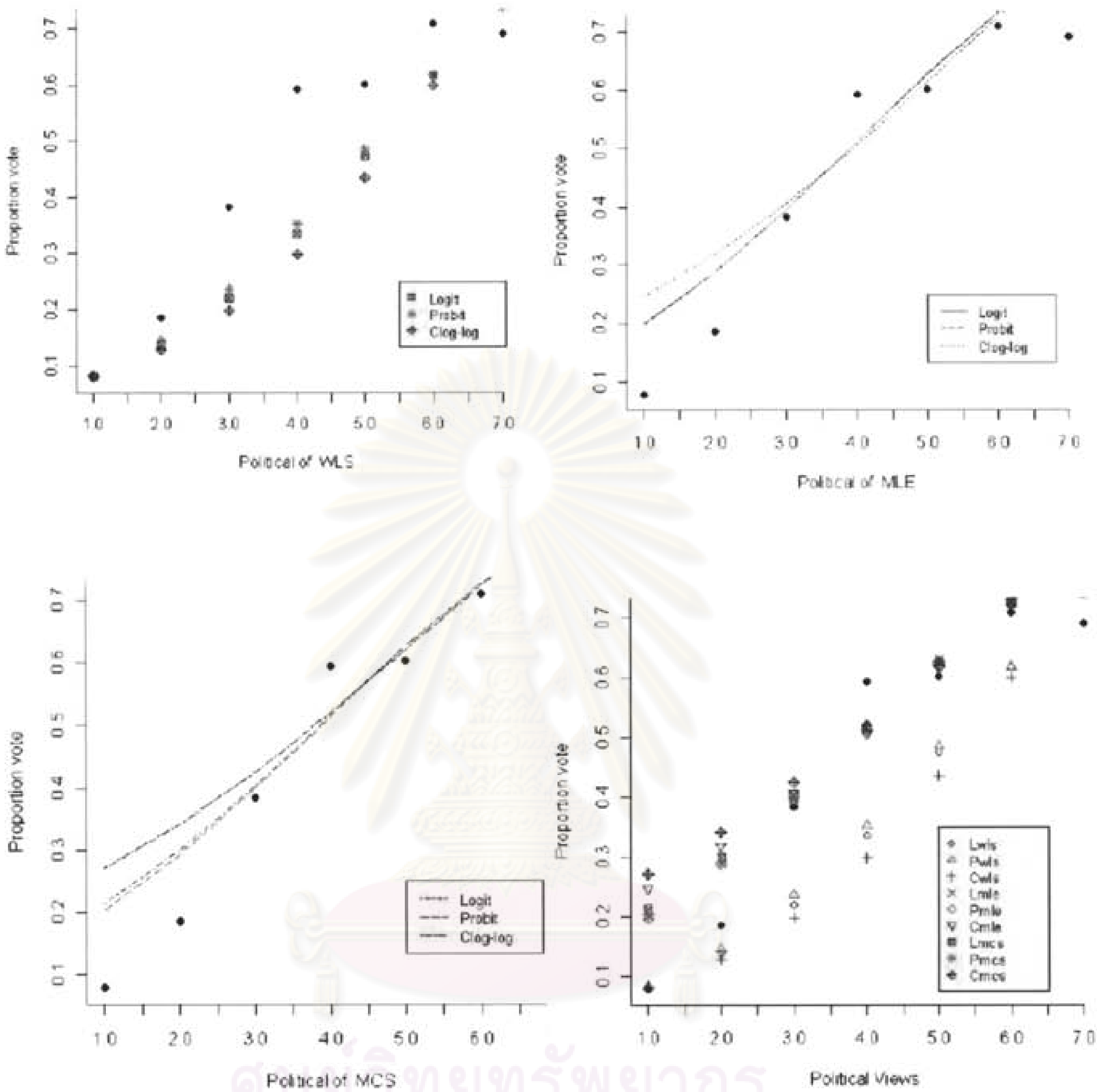
วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	104.77180	0.00000	89.33742	0.00000	145.63180	0.00000
MLE	15.59553**	0.00810	15.76384	0.00755	20.36838	0.00107
MCS	15.75094*	0.00759	15.79503	0.00745	20.97647	0.00082

ข้อมูลการสำรวจทางสังคม ปี 1982 ของคนผิวขาว ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่อง Political Views เพื่อดูว่าคนผิวขาวจะเลือก Reagan เป็นประธานาธิบดี ข้อมูลจาก Clogg and Shockey (1988) Political Views ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ การเลือก หรือ ไม่เลือก Reagan เป็นประธานาธิบดี ข้อมูล คือ จำนวนความถี่ของการบุคคลที่เลือก Reagan และไม่เลือก Reagan ตาม Political Views ข้อมูลจากพบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS)วิธีภาวะน่าจะเป็นสูงสุด

(MLE) และวิธีไคกำลังสองต่ำสุด (MCS) จำแนกตาม Political Views 7 ระดับ แสดงไว้ในตารางที่ 4.15

การทดสอบความเหมาะสมของตัวแบบ ภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.16 โดยค่าสถิติ Deviance ของตัวแบบโลจิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 15.59553 ให้ค่า P-value เท่ากับ 0.00810 น้อยกว่าระดับนัยสำคัญ 0.05 จากข้อมูลชุดนี้ไม่ควรเลือกตัวแบบโลจิท ภายใต้วิธี MLE ในการวิเคราะห์ข้อมูล ถึงแม้ว่าจะให้ค่าสถิติ Deviance น้อยที่สุดจากทุกกรณีก็ตาม

จากรูปที่ 4.8 แสดงการพล็อตความน่าจะเป็นผู้ที่ทำการเลือก Reagan เป็นประธานาธิบดี ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อ Political Views ต่างกัน ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกัน ดังรูปที่ 4.8 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตค่อนข้างเป็นเส้นตรง การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ห่างกับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Deviane มีค่ามาก กล่าวได้ว่าตัวแบบโลจิท ที่ประมาณได้ด้วยวิธี MLE ยังไม่มีความเหมาะสมกับข้อมูลชุดที่ 8



รูปที่ 4.8 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 8

ตารางที่ 4.17 ข้อมูลปีของการศึกษาและค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี

Year of Edication	ขนาดตัวอย่าง	ค่าตอบสนอง	ค่าประมาณจากตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์แต่ละวิธี								
			Logit			Probit			C log- log		
			WLS	MLE	MCS	WLS	MLE	MCS	WLS	MLE	MCS
3	16	12	13.0	13.8	13.7	12.9	13.8	13.8	13.8	14.6	14.7
4	20	15	15.1	16.5	16.4	15.0	16.5	16.4	15.5	17.3	17.4
5	41	27	28.3	32.0	31.7	28.3	31.8	31.7	27.8	32.9	33.1
6	56	42	34.4	40.6	40.3	34.8	40.3	40.1	32.1	41.0	41.2
7	84	53	44.8	55.8	55.3	46.2	55.3	55.1	39.8	55.1	55.5
8	251	166	112.9	149.9	148.7	119.7	148.6	148.1	96.4	145.5	146.7
9	123	59	45.4	64.7	64.3	49.7	64.3	64.2	37.6	62.1	62.7
10	199	87	58.7	90.3	90.0	66.6	90.3	90.1	47.9	86.5	87.4
11	207	86	47.7	79.4	79.4	55.9	79.8	79.8	38.8	76.6	77.5
12	953	305	168.2	302.8	304.2	202.8	305.7	306.3	138.1	298.0	301.7
13	210	48	27.9	54.3	54.8	34.3	54.8	55.1	23.4	55.1	55.8
14	206	46	20.4	42.6	43.3	25.1	42.8	43.1	17.5	45.0	45.7
15	73	16	5.3	11.9	12.2	6.5	11.8	11.9	4.7	13.2	13.4
16	253	28	13.5	32.3	33.2	15.8	30.9	31.4	12.5	37.9	38.5
17	63	6	2.4	6.2	6.4	2.7	5.7	5.8	2.4	7.8	7.9
18	50	1	1.4	3.8	4.0	1.4	3.3	3.3	1.4	5.1	5.1

องศาความเป็นอิสระ df = 14

ตารางที่4.18 การตรวจสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ในข้อมูลชุดที่ 9

วิธีการประมาณค่าพารามิเตอร์	ความเหมาะสมของตัวแบบ					
	Logit		Probit		C log-log	
	Deviance	P-value	Deviance	P-value	Deviance	P-value
WLS	301.95510	0.00000	178.28250	0.00000	490.09900	0.00000
MLE	19.08540	0.16171	18.42963**	0.18791	31.15906	0.00527
MCS	19.25154	0.15556	18.46386*	0.18646	31.34754	0.00495

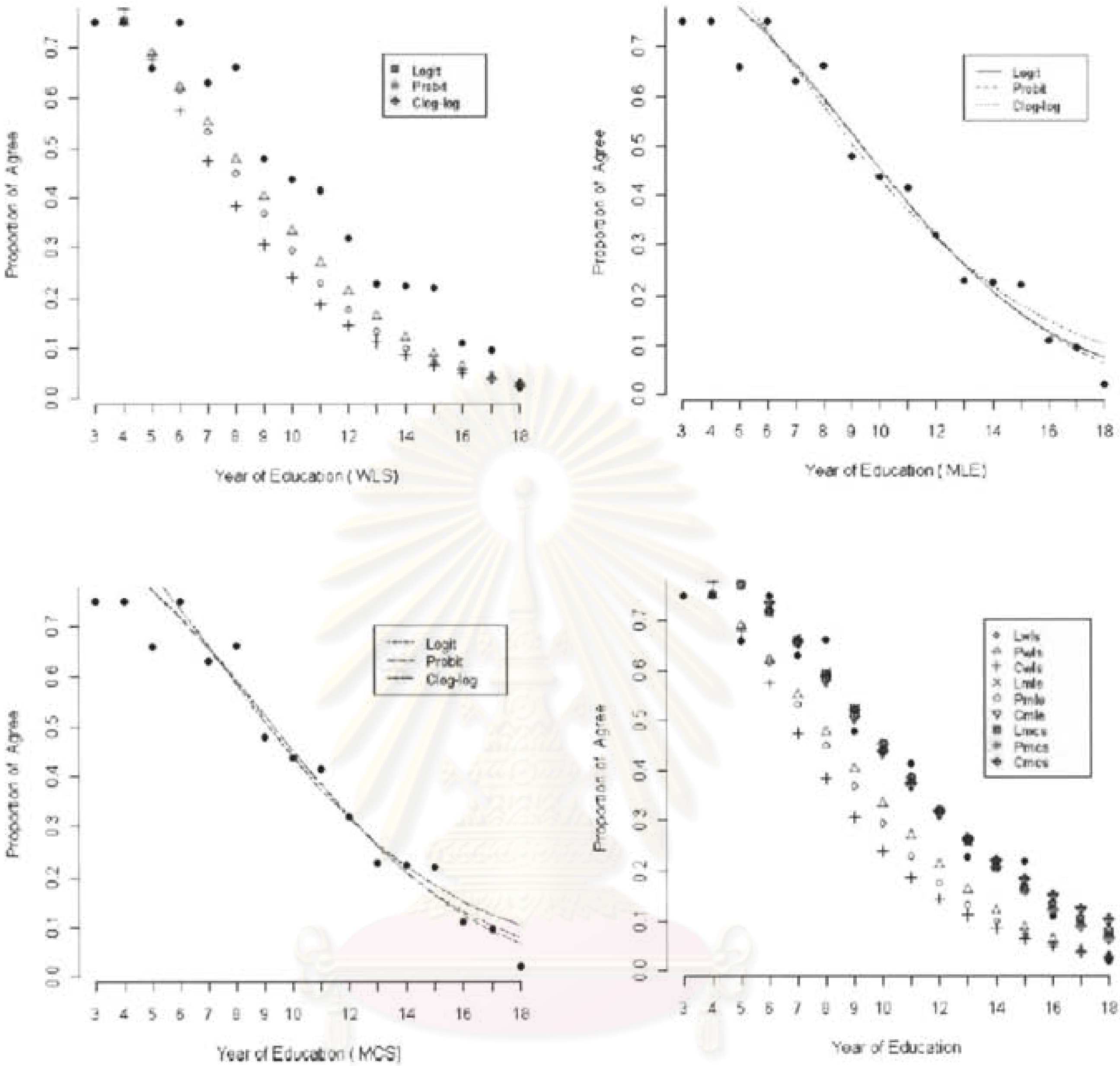
ข้อมูลการสำรวจความคิดเห็นเกี่ยวกับบทบาทของผู้หญิงที่มีต่อสังคม ทำการสำรวจทั้ง เพศหญิงและเพศชาย ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องปีของการสำเร็จการศึกษา ต่อความคิดเห็นของการเห็นด้วยหรือไม่เห็นด้วยของ คำกล่าวที่ว่า “ผู้หญิงมีหน้าที่ดูแลบ้านและอนุญาตให้ทำงานนอกบ้านได้เหมือนผู้ชาย” ข้อมูลจาก Haberman(1978) ปีการศึกษา (X) ซึ่งเป็นตัวประกอบหนึ่งของตัวแปรอธิบาย ส่วนตัวแปรตอบสนอง คือ เห็นด้วยและหรือไม่เห็นด้วยเมื่อปีการศึกษาต่างระดับกัน ข้อมูล คือ จำนวนความถี่ของผู้ที่เห็นด้วยและผู้ที่ไม่เห็นด้วย ณ ปีการศึกษาที่ระดับต่าง ๆ พบว่าค่าประมาณที่ได้จากตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์ที่ศึกษา คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) จำแนกตามระดับปีที่ศึกษา แสดงไว้ในตารางที่ 4.17

การทดสอบความเหมาะสมของตัวแบบภายใต้วิธีการประมาณค่าพารามิเตอร์ ดังตารางที่ 4.18 โดยค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธี MLE ให้ค่า Deviance เท่ากับ 18.42963 ให้ค่า P-value เท่ากับ 0.18791 ที่องศาความเป็นอิสระเท่ากับ 14 ความน่าจะเป็นในการยอมรับตัวแบบโพรบิท มีมากกว่าความน่าจะเป็นตัวแบบโลจิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก จากข้อมูลด้านการศึกษาชุดนี้ ควรเลือกตัวแบบโพรบิท ภายใต้วิธี MLE เพราะให้ค่าสถิติ Deviance น้อยที่สุด จากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ จึงถือได้ว่าตัวแบบโพรบิท ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดนี้มากที่สุด และให้ค่าประมาณที่ได้จากตัวแบบใกล้เคียงกับค่าสังเกตจากข้อมูลจริงมากที่สุด ภายใต้ระดับนัยสำคัญ 0.05

ในขณะที่ วิธี WLS ของทุกตัวแบบให้ค่า Deviance สูงมาก อาจเป็นเพราะเหตุผลดังนี้ เนื่องจากข้อมูลชุดที่ 9 นี้ มีจำนวนเซลล์มาก และความถี่ในแต่ละเซลล์ก็ค่อนข้างน้อย และถ้าตัวแปรอธิบายเป็นแบบต่อเนื่อง วิธี WLS อาจไม่เหมาะสม เนื่องจากอาจมีเพียง 1 ค่าสังเกตในแต่ละ setting นั้น จึงทำให้มีผลต่อการประมาณด้วยวิธี WLS แต่ข้อมูลในลักษณะนี้ไม่มีปัญหากับวิธี MLE ถึงแม้ว่าในเซลล์นั้นจะมีค่าเป็น 0 วิธี MLE ก็สามารถทำได้ อาจแทนค่าคงที่ๆมีค่าน้อยมากในเซลล์เพื่อให้สามารถคำนวณค่าประมาณจากวิธี MLE พร้อมกับใช้วิธีการย้อนซ้ำปรับค่าถ่วงน้ำหนักต่างๆ ต่างกับส่วนของวิธี WLS โดยตรงที่ เมื่อแทนค่าในเซลล์ศูนย์ ด้วยค่าเช่น 0.5 หรือน้อยกว่านี้อีกมากๆ อาจทำให้เทอมความแปรปรวนมีค่าสูงหรือต่ำผิดกว่าปกติ จนกระทั่งมีผลกระทบอย่างมากต่อการวิเคราะห์การถ่วงน้ำหนัก ทำให้ลดความน่าเชื่อถือถือทั้งผลลัพธ์และข้อมูล กรณีเช่นนี้ควรใช้วิธี MLE

จากรูปที่ 4.9 แสดงการพล็อตความน่าจะเป็นของการเห็นด้วยที่จะเห็นผู้หญิงมีบทบาทในการทำงานนอกบ้าน ตามตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อระดับการศึกษาที่สูงขึ้น ทำการประมาณค่าด้วยวิธีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด(MLE) และ วิธีไคกำลังสองต่ำสุด (MCS)

และทำการพล็อตความน่าจะเป็นรวมทั้ง 3 ตัวแบบด้วยวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ในกราฟเดียวกันดังรูปที่ 4.9 ดูจากลักษณะของกราฟความน่าจะเป็นในการพล็อตเป็นรูปตัวเอสแบบกลับข้าง การกระจายของความน่าจะเป็นที่เกิดจากการพยากรณ์ ใกล้กับจุดความน่าจะเป็นของค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงเป็นเหตุให้ค่าสถิติ Deviane มีค่าน้อย ยกเว้นกรณี WLS ที่มีการกระจายความน่าจะเป็นจากการพยากรณ์ห่างจากจุดความน่าจะเป็นจากค่าตอบสนองที่สังเกตได้ค่อนข้างมาก จึงทำให้ตัวแบบทั้ง 3 ตัวแบบที่ประมาณด้วยวิธี WLS ไม่มีความเหมาะสมกับข้อมูลชุดที่ 9 นี้ แต่การประมาณด้วยวิธี MLE และ MCS การกระจายของความน่าจะเป็นค่อนข้างใกล้เคียงกับค่าความน่าจะเป็นของค่าตอบสนอง กล่าวได้ว่าตัวแบบโพรบิท ที่ประมาณได้ด้วยวิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 9



รูปที่ 4.9 การพล็อตความน่าจะเป็นของตัวแบบทั้ง 3 ตัวแบบ ด้วยวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ของข้อมูลชุดที่ 9

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบ วิธีการประมาณค่าพารามิเตอร์ในตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ซึ่งวิธีการประมาณค่าพารามิเตอร์ที่ใช้ในงานวิจัยครั้งนี้ คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS) วิธีภาวะน่าจะเป็นสูงสุด (MLE) และ วิธีไคกำลังสองต่ำสุด (MCS) โดยที่ตัวแปรตอบสนอง (Y) เป็นตัวแปรเชิงคุณภาพมี 2 ค่า คือ 0 หรือ 1 และตัวแปรอธิบาย(X) 1 ตัวแปร การเปรียบเทียบกระทำภายใต้ข้อมูล 9 ชุด เป็นผลงานวิจัยการทดลองของ Draper(1972), Ashford(1970), Cornfield (1962), Martin(1942), Muhammad(1990), Strand(1930), Montgomery(1982), Clogg(1988) และ Haberman(1978) ตัวอย่างข้อมูลส่วนใหญ่ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ยกเว้นตัวอย่างชุดที่ 1 ใช้ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ตัวอย่างชุดที่ 4 และ 6 ใช้ตัวแบบโพรบิทในการวิเคราะห์ ข้อมูลที่ใช้ในงานวิจัยครั้งนี้ตัวอย่างชุดที่ 1-3 เป็นข้อมูลด้านการแพทย์ ตัวอย่างชุดที่ 4-6 เป็นข้อมูลทางด้านวิทยาศาสตร์(ชีววิทยา) ตัวอย่างชุดที่ 7 เป็นข้อมูลทางด้านวิศวกรรมศาสตร์ และตัวอย่างที่ 8-9 เป็นข้อมูลทางด้านสังคมศาสตร์ ข้อมูลที่วิธีการวิเคราะห์ข้อมูล คือ การประมาณค่าด้วยตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก โดยใช้ตัวสถิติ Deviance เป็นเกณฑ์ในการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ และ เลือกตัวแบบ ให้เหมาะสมกับลักษณะของข้อมูล ผลการวิจัยสรุปได้ ดังนี้

1. ข้อมูลตัวอย่างชุดที่ 1 เป็นข้อมูลทางด้านการแพทย์ ของ Draper(1972) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบคอมพลีเมนทารีล็อก-ล็อก ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโลจิทภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 13.76141 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ให้ค่า p-value มีค่าเท่ากับ 0.017198 น้อยกว่าระดับนัยสำคัญ 0.05 สรุปคือไม่มีตัวแบบใดเหมาะสมกับข้อมูลชุดที่ 1

2. ข้อมูลตัวอย่างชุดที่ 2 เป็นข้อมูลทางด้านการแพทย์ของ Ashford(1970) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 4.05551 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance มีค่าเท่ากับ 4.05551 ให้ค่า p-value

เท่ากับ 0.77336 ผลลัพธ์คือ การยอมรับตัวแบบโพรบิทภายใต้วิธี MLE มีมากว่าตัวแบบโลจิทและตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบโพรบิท ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 2 มากที่สุด

3. ข้อมูลตัวอย่างชุดที่ 3 เป็นข้อมูลทางด้านการแพทย์ของ Cornfield (1962) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 5.88427 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance มีค่าเท่ากับ 5.88427 ให้ค่า p-value เท่ากับ 0.43628 ผลลัพธ์คือการยอมรับตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE มีมากกว่าตัวแบบโลจิทและตัวแบบโพรบิท ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 3 มากที่สุด

4. ข้อมูลตัวอย่างชุดที่ 4 เป็นข้อมูลทางด้านวิทยาศาสตร์(ชีววิทยา) ของ Martin(1942) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโพรบิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโลจิท ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 1.34806 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance เท่ากับ 1.34806 ให้ค่า p-value เท่ากับ 0.71775 ผลลัพธ์คือการยอมรับตัวแบบโลจิท ภายใต้วิธี MLE มีมากกว่าตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบโลจิท ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 4 มากที่สุด

5. ข้อมูลตัวอย่างชุดที่ 5 เป็นข้อมูลทางด้านวิทยาศาสตร์(ชีววิทยา) ของ Muhammad(1990) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 0.27191 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance เท่ากับ 0.27191 ให้ค่า p-value เท่ากับ 0.96522 ผลลัพธ์คือ การยอมรับตัวแบบโพรบิทภายใต้วิธี MLE มีมากว่าตัวแบบโลจิทและตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบโพรบิท ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 5 มากที่สุด

6. ข้อมูลตัวอย่างชุดที่ 6 เป็นข้อมูลทางด้านวิทยาศาสตร์(ชีววิทยา) ของ Strand(1930) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโพรบิท ผลปรากฏว่าพบว่าค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 19.12758 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance

เท่ากับ19.12758 ให้ค่า p-value เท่ากับ 0.00395 น้อยกว่าระดับนัยสำคัญ 0.05 สรุปคือไม่มีตัวแบบใดเหมาะสมกับข้อมูลชุดที่ 6

7. ข้อมูลตัวอย่างชุดที่ 7 เป็นข้อมูลทางด้านวิศวกรรมศาสตร์ ของ Montgomery(1982)ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าพบว่า ค่าสถิติ Deviance ของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 0.40519 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance เท่ากับ 0.40519 และค่า p-value เท่ากับ 0.99994 ผลลัพธ์คือ การยอมรับตัวแบบคอมพลีเมนทารีล็อก-ล็อกภายใต้วิธี MLE มีมากกว่าตัวแบบโลจิทและตัวแบบโพรบิท ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 7 มากที่สุด

8. ข้อมูลตัวอย่างชุดที่ 8 เป็นข้อมูลทางด้านสังคมศาสตร์ ของ Clogg (1988) ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโลจิทภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 15.59553 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance เท่ากับ 15.59553 ให้ค่า p-value เท่ากับ 0.00395 น้อยกว่าระดับนัยสำคัญ 0.05 สรุปคือไม่มีตัวแบบใดเหมาะสมกับข้อมูลชุดที่ 8

9. ข้อมูลตัวอย่างชุดที่ 9 เป็นข้อมูลทางด้านสังคมศาสตร์ ของ Haberman(1978)ได้ใช้เทคนิคการวิเคราะห์ด้วยตัวแบบโลจิท ผลปรากฏว่าค่าสถิติ Deviance ของตัวแบบโพรบิท ภายใต้วิธีการประมาณค่าพารามิเตอร์แบบภาวะน่าจะเป็นสูงสุด (MLE) ให้ค่า Deviance = 18.42963 น้อยที่สุดจากทุกกรณีของวิธีการประมาณค่าพารามิเตอร์ ตัวสถิติ Deviance เท่ากับ 18.42963 ให้ค่า p-value เท่ากับ 0.18791 ผลลัพธ์คือ การยอมรับตัวแบบโพรบิทภายใต้วิธี MLE มีมากว่าตัวแบบโลจิทและตัวแบบคอมพลีเมนทารีล็อก-ล็อก ภายใต้วิธี MCS,WLE กล่าวได้ว่าตัวแบบโพรบิท ภายใต้วิธี MLE มีความเหมาะสมกับข้อมูลชุดที่ 9 มากที่สุด

สรุปผลการวิจัยเกี่ยวกับวิธีการประมาณค่าพารามิเตอร์ที่ใช้ในงานวิจัยครั้งนี้ คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS), วิธีภาวะน่าจะเป็นสูงสุด (MLE),และ วิธีไคกำลังสองต่ำสุด (MCS) ผลปรากฏว่าวิธี MLE เป็นวิธีการประมาณค่าพารามิเตอร์ที่ดีที่สุดเนื่องจากว่า วิธี MLE มีวิธีการย้อนซ้ำปรับค่าถ่วงน้ำหนักต่างๆ ในแต่ละรอบ แต่วิธี MCS และ WLS ไม่มีการย้อนซ้ำ ถ้าเซลล์ใดมีค่าสังเกตเป็น 0 วิธี MCS และ WLS ไม่สามารถคำนวณหาค่าประมาณได้ แต่วิธี MLE สามารถทำได้ ต่างกับส่วนของวิธี WLS โดยตรงที่ เมื่อแทนค่าในเซลล์ศูนย์ ด้วยค่าเช่น 0.5 หรือน้อยกว่านี้อีกมากๆ อาจทำให้เทอมความแปรปรวนมีค่าสูงหรือต่ำผิดกว่าปกติ จนกระทั่งมี

ผลกระทบอย่างมากต่อการวิเคราะห์การถ่วงน้ำหนัก ทำให้ลดความน่าเชื่อถือถือทั้งผลลัพธ์และข้อมูล กรณีเช่นนี้ควรใช้วิธี MLE

สรุปผลการวิจัยเกี่ยวกับตัวแบบ ซึ่งตัวแบบที่ใช้ในงานวิจัยครั้งนี้ คือ ตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก หลังจากทำการปรับค่าตัวแปรอธิบายแล้ว ผลที่ได้คือลิงค์ฟังก์ชันเปลี่ยนไป เพราะว่า ขนาดความแปรปรวนขึ้นอยู่กับขนาดของข้อมูล และความแตกต่างของค่าพารามิเตอร์ที่ประมาณได้จะขึ้นอยู่กับขนาดของ σ เท่านั้น ดังนั้นจะเห็นได้ว่าตัวแปรอธิบายมีผลต่อการเลือกลิงค์ฟังก์ชัน

การพิจารณาเลือกลิงค์ฟังก์ชัน

1. เมื่อตัวแปรตอบสนองเป็นแบบ 2 กลุ่ม และตัวแปรอธิบาย เป็นตัวแปรเชิงกลุ่ม ควรเลือกลิงค์ฟังก์ชัน ของโลจิท โพรบิท หรือ ล็อกลิเนียร์
2. ถ้าตัวแปรตอบสนองเป็นแบบ 2 กลุ่ม และตัวแปรอธิบาย เป็นตัวแปรแบบเนื่อง ควรเลือกลิงค์ฟังก์ชัน ของโลจิท หรือ โพรบิท
3. ตัวแปรตอบสนองเป็นข้อมูลระยะยาว (Censored duration data) และตัวแปรอธิบายเป็นทั้งเชิงกลุ่มหรือต่อเนื่อง ควรเลือกลิงค์ฟังก์ชันของ ล็อกลิเนียร์ โลจิท หรือ คอมพลีเมนทารีล็อก-ล็อก

5.2 อธิปรายผล

จากการวิจัยที่ทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ 3 วิธี คือ วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (WLS), วิธีภาวะน่าจะเป็นสูงสุด (MLE), และ วิธีไคกำลังสองต่ำสุด (MCS) โดยที่ตัวแปรตอบสนอง (Y) เป็นตัวแปรเชิงคุณภาพมี 2 ค่า คือ 0 หรือ 1 พบว่าตัวอย่างทั้ง 9 ชุด มีตัวแบบที่เหมาะสมกับลักษณะของแต่ละข้อมูล ภายใต้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดให้ค่าสถิติ Deviance ต่ำสุดทุกกรณี ซึ่งผลที่ได้ สอดคล้องกับงานวิจัยของกาญจนา พานิชการ(2539) และ Huhn,M(2000)

5.3 ข้อเสนอแนะ

1. งานวิจัยครั้งนี้เป็นการศึกษาและเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์แบบจุดของตัวประมาณพารามิเตอร์ในตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก เท่านั้น ซึ่งเป็นที่น่าสนใจที่ทำการศึกษาถึงวิธีการประมาณช่วงความเชื่อมั่นสำหรับตัวประมาณค่าพารามิเตอร์ในตัวแบบทั้ง 3 ตัวแบบ

2. ทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์แบบจุดในตัวแบบโลจิท ตัวแบบโพรบิท และตัวแบบคอมพลีเมนทารีล็อก-ล็อก เมื่อตัวแปรตอบสนองมีลักษณะเป็นแบบมีลำดับ (Ordinal response)

3. ประมาณค่าพารามิเตอร์ในตัวแบบทั้ง 3 ตัวแบบด้วยวิธีเบส์ และวิธีโมเมนต์ เทียบกับวิธีภาวะน่าจะเป็นสูงสุด

4. ในการวิเคราะห์ข้อมูลนอกจากการใช้ตัวแบบ 3 ตัวแบบแล้ว อาจทำการวางทดลอง โดยใช้วิธีการสถิติที่เรียกว่า การวิเคราะห์ความแปรปรวน ซึ่งใช้วิธีการทดสอบแบบบล็อกสุ่มสมบูรณ์ (Randomized Complete Block Design : RCBD) เพื่อทดสอบว่าตัวแปรอธิบายมีผลต่อเหตุการณที่เราสนใจหรือไม่ โดยกำหนดให้ตัวแปรอธิบายเป็นหน่วยทดลองสามารถจำแนกออกเป็นกลุ่ม ซึ่งเรียกว่า บล็อกได้

5. ในการวิเคราะห์ข้อมูลพิจารณาใช้ตัวแบบโลจิท ตัวแบบโพรบิท และ ตัวแบบคอมพลีเมนทารีล็อก-ล็อก เมื่อตัวแปรตอบสนองมี 2 กลุ่ม กับตัวแปรอธิบาย ในการวิเคราะห์ข้อมูลชุดที่ 1 ชุดที่ 6 และ ชุดที่ 8 ถ้าทำการวิเคราะห์ด้วยตัวแบบทั้ง 3 ตัวแบบไม่ได้ ควรจะทำการวิเคราะห์ด้วยตัวแบบล็อกลิเนียร์ เพราะตัวแบบล็อกลิเนียร์สามารถวิเคราะห์ความสัมพันธ์ของตัวแปรครั้งละหลายตัวแปร ตัวแปรละหลายระดับ และยังเป็นตัวแบบที่สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรเชิงกลุ่มได้ดี

รายการอ้างอิง

ภาษาไทย

กาญจนา พานิชการ. 2539. การประมาณค่าพารามิเตอร์ในสมการถดถอยโลจิสติคด้วยภาวะน่าจะเป็นสูงสุดและฟังก์ชันจำแนกประเภท. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

ชนิศวรา ฉัตรแก้ว. 2543. การวิเคราะห์การถดถอยเมื่อตัวแปรตามมีสองลักษณะ โดยใช้ตัวแบบความน่าจะเป็นเชิงเส้น ตัวแบบโพรบิท และตัวแบบโลจิท . วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

ทัศนาพร จงเกตุจรณ์. 2546. การประมาณค่าพารามิเตอร์ของตัวแบบถดถอยโลจิสติคทวินาม. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

พิมลรัตน์ รัตนเพชร. 2547. การวิเคราะห์การถดถอยที่ตัวแปรตามมีค่าเป็น 2 ลักษณะในกรณีที่มีการแปลงข้อมูล. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

วีรานันท์ พงศาภักดี. 2544. การวิเคราะห์ข้อมูลเชิงกลุ่ม ทฤษฎีและการประยุกต์ (กับGLIM และ SPSS/FW). พิมพ์ครั้งที่ 2.. นครปฐม : โรงพิมพ์มหาวิทยาลัยศิลปากร,

เรวดี เรืองอยู่. 2547. การเปรียบเทียบการประมาณค่าพารามิเตอร์ด้วยวิธีแบบบริดจ์ ภาวะน่าจะเป็นสูงสุด และฟังก์จำแนกประเภทในสมการถดถอยโลจิสติค. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

ภาษาอังกฤษ

Agresti, A. 1990. Categorical Data Analysis. New York : Wiley,

Amemiya, T. 1974. Bivariate probit analysis: Minimum chi-square method. Journal of the American Statistical Association 69 : 940-944.

Ashford,J.R. and Sowden, R.R. 1970. Multi-variate probit analysis. Biometrics 26 : 535-46.

Berkson, J. 1944. Application of the logistic function to bio–assay. Journal of the American Statistical Association 39 : 357-365.

Berkson, J. 1955. Maximum likelihood and minimum chi-square of the logistic function. Journal of the American Statistical Association 50 : 130-162.

Berkson, J. 1955. Estimation of the Integrated normal curve by minimum normit chi-square with particular reference to bio-assay. Journal of the American Statistical Association 50 : 529-549.

Berkson, J. 1980. Minimum chi-square, not maximum likelihood. Annals of Statistics 8 : 457-487

Bliss,C.I. 1935. The Calculation of the Dosage- mortality Curve . Ann.Appl. Biol 22 : 134-167.

Collett, D. 2003. Modelling Binary Data. 2nd ed. London : Chapman and Hall/CRC,

Clogg,C.C., and J. W. Shockey. 1988. Multivariate analysis of discrete data In Handbook of Multivariate Experimental Psychology, ed. By J. R. Nesselroade and R.B Cattell. New York : Plenum Press,

Dobson,A.J. 2001. An Introduction to Generalized Linear Models. London : Chapman and Hall,

Draper,C. C., Voller, A. and Carpenter, R. G. 1972. The epidemiologic interpretation of serologic data in malaria. American Journal of Tropical Medicine and Hygiene 26 : 696-703.

Fahrmeir , L. and G.Tutz. 1994. Multivariate Statistical Modelling Based on Generalized Linear models. New York : Springer – Verlag,

Finney, D.J. 1971. Probit Analysis. 3rd ed. Cambridge University Press

Griffiths, W.E., R.C. Hill and G.G. Judge. 1993. Learning and Practicing Econometri. New York : Jonh Wiley and Sons,

Harris,R.R & Kanji G.K. 1983. On the Use of Minimum Chi-Square Estimation. Journal of Royal Statistical Society 32 : 379-394.

Huhn, M.. 2000. Maximum likelihood vs. minimum chi-square—A general comparison with applicatio to the estimation of recombination fractions in two-point linkage analysis. NCR Canada Genome 43 : 853-856.

Horowitz.L.J ; Savin. N.E. 2001. Binary Response Models : Logit, Probit and Semiparametrics. The Journal of Economic Perspectives 15 No. 4 : 43-56.

Martin,J.T. 1942. The problem of the evaluation of rotenone containing plants. vi The toxicity of l-elliptone and of poisons applied jointly, with further observations on the rotenone equivalent method of assessing the toxicity of derris root. Ann.Appl.Bio 29 : 69-81.

McCullagh ,P., & J.A. Nelder . 1989. Generalized Linear Models. 2nd ed. New York : Chapman and Hall,

Montgomery, D.C. and Peck, E.A. 1982. Introduction to linear regression analysis.New York : Wiley,

Muhammad. F & Khan, A & Ahmad. 1990. Logistic Regression Analysis in Dose Response Studies. Journal of Islamic Academy of Science 3:2 : 103-106.

Nagler.J . 1994. An Alternative Estimator to Logit and Probit.. Journal of Political Science : 230-255.

Nelder, J.A. & Wedderburn, R.W.M. 1972. Generalized Linear Models . J.Roy Statest.Soc.Ser A135 : 370-384.

Power, D. A. and Xie, Y. 1999. Statistical Methods for Categorical Data Analysis.San Diego : Academic Press,

Seppo Laaksonen . 2006. Alternative Link Functions in Survey Estimation Under Missingness. Proceedings of Q2006 European Conference on Quality in Survey Statistics .

Strand, A.L. 1930. Measuring the toxicity of insect fumigants. Industr. Engng Chem 2 : 4-8

ภาคผนวก

ภาคผนวก ก

ตัวอย่างการเขียนโปรแกรม

ข้อมูลทางด้านวิศวกรรมศาสตร์

WEIGHTED LEAST SQUARE

```
table.S.1 <-data.frame(load=c(2500,2700,2900,3100,3300,3500,3700,3900,4100,4300),
        n=c(50,70,100,60,40,85,90,50,80,65),
        y=c(10,17,30,21,18,43,54,33,60,51))
log.load<- log(table.S.1$load)
p<-c(table.S.1$y/table.S.1$n)
q<- 1-p
Var<-c(table.S.1$n*p*q)
Vi<-1/Var
W<-diag(Var)
V<-1/W
InvW<-solve(W)
 nuhat.Lwls<- log(p/q)
nuhat.Pwls<-qnorm(p)
nuhat.Cwls<-log(-log(q))
X.wls<-as.matrix(cbind(rep(1,10), log.load))
Covb.wls<-solve(t(X.wls)%*%InvW%*%X.wls)
b.Lwls<-as.matrix(Covb.wls%*%t(X.wls)%*%InvW%*%nuhat.Lwls)
b.Pwls<-as.matrix(Covb.wls%*%t(X.wls)%*%InvW%*%nuhat.Pwls)
b.Cwls<-as.matrix(Covb.wls%*%t(X.wls)%*%InvW%*%nuhat.Cwls)
yhat.Lwls<-as.matrix(X.wls%*%b.Lwls)
yhat.Pwls<-as.matrix(X.wls%*%b.Pwls)
yhat.Cwls<-as.matrix(X.wls%*%b.Cwls)
prob.Lwls<-(exp(yhat.Lwls))/(1+exp(yhat.Lwls))
prob.Pwls<-pnorm(yhat.Pwls)
```

```
prob.Cwls<-1-exp(-exp(yhat.Cwls))
ypred.Lwls<-c(table.S.1$n*prob.Lwls)
ypred.Pwls<-c(table.S.1$n*prob.Pwls)
ypred.Cwls<-c(table.S.1$n*prob.Cwls)
D.Lwls<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Lwls)+(table.S.1$n-table.S.1$y)*
        log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Lwls))))   #deviance of  logit
D.Pwls<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Pwls)+(table.S.1$n-table.S.1$y)*
        log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Pwls))))#deviance of  probit
 D.Cwls<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Cwls)+(table.S.1$n-table.S.1$y)*
        log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Cwls))))#deviance of  cloglog
P.D.Lwls<-1-pchisq(D.Lwls,df=8)
P.D.Pwls<-1-pchisq(D.Pwls,df=8)
P.D.Cwls<-1-pchisq(D.Cwls,df=8)
plot(log.load, table.S.1$y/table.S.1$n, pch=16, xlab="Log.Load of  WLS",
        ylab=" Proportion of fasteners that fail ", bty="L", axes=F)
axis(1, at=seq(7,9, .1))
axis(2, at=seq(0, 1, .1))
points(log.load, prob.Lwls,pch = 1,col=2)
points (log.load, prob.Pwls,pch = 2,col=3)
points (log.load, prob.Cwls,pch = 3,col=4)
legend(x=8.3, y=.35, legend=c("Logit","Probit","Clog-log"),pch=c(1,2,3),col=c(2,3,4), cex=.85,
        text.width=1, adj =c(0,.5))

                    #MAXIMUM LIKELIHOOD ESTIMATION#

fit.logit<-glm(table.S.1$y/table.S.1$n~log.load,weights=table.S.1$n,
        family=binomial(link=logit))
fit.probit<-glm(table.S.1$y/table.S.1$n~log.load,weights=table.S.1$n,
        family=binomial(link=probit))
fit.cloglog<-glm(table.S.1$y/table.S.1$n~log.load, weights=table.S.1$n,
        family=binomial(link=cloglog))
P.D.Lmle<-1-pchisq(fit.logit$deviance,df=8)
```

```
P.D.Pmle<-1-pchisq(fit.probit$deviance,df=8)
P.D.Cmle<-1-pchisq(fit.cloglog$deviance,df=8)
plot(log.load,table.S.1$y/table.S.1$n, pch=16, xlab="Log.Load of  MLE",
        ylab=" Proportion of fasteners that faill", bty="L", axes=F)
 axis(1, at=seq(7,9, .1))
axis(2, at=seq(0, 1, .1))
lines(log.load, fitted(fit.logit),lty = 1,col=2)
lines(log.load, fitted(fit.probit),lty = 2,col=3)
lines(log.load, fitted(fit.cloglog),lty = 3,col=4)
 legend(x=8.25, y=.35, legend=c("Logit","Probit","Clog-log"),
          lty=c(1,2,3), col=c(2,3,4),cex=.85, text.width=1, adj =c(0,.5))

                              #MINIMUM CHI-SQUARE#

 p<-c(table.S.1$y/table.S.1$n)
 q<- 1-p
 z.Lmcs<- log(p/q)
 z.Pmcs<- qnorm(p)
z.Cmcs<- log(-log(q))
 W.1<-c(table.S.1$n*p*q)  # weights of logit
 W.2<-c((table.S.1$n*(dnorm(z.Pmcs)^2))/(p*q))# weights of probit
 A<-c((1/(-log(q)*table.S.1$n*q)))
 W.3<-c((1/((A^2)*table.S.1$n*p*q)))# weights of cloglog
 W.Lmcs<-diag(W.1)
 W.Pmcs<-diag(W.2)
 W.Cmcs<-diag(W.3)
 X<-as.matrix(cbind(rep(1,10),log.load))
 Covb.Lmcs<-solve(t(X)%*%W.Lmcs%*%X)
 Covb.Pmcs<-solve(t(X)%*%W.Pmcs%*%X)
 Covb.Cmcs<-solve(t(X)%*%W.Cmcs%*%X)
 b.Lmcs<-as.matrix(Covb.Lmcs%*%t(X)%*%W.Lmcs%*%z.Lmcs)
 b.Pmcs<-as.matrix(Covb.Pmcs%*%t(X)%*%W.Pmcs%*%z.Pmcs)
```

```
b.Cmcs<-as.matrix(Covb.Cmcs%*%t(X)%*%W.Cmcs%*%z.Cmcs)
yhat.Lmcs<-as.matrix(X%*%b.Lmcs)
yhat.Pmcs<-as.matrix(X%*%b.Pmcs)
yhat.Cmcs<-as.matrix(X%*%b.Cmcs)
prob.Lmcs<-exp(yhat.Lmcs)/(1+exp(yhat.Lmcs))
prob.Pmcs<-pnorm(yhat.Pmcs)
prob.Cmcs<-1-exp(-exp(yhat.Cmcs))
ypred.Lmcs<-c(table.S.1$n*prob.Lmcs)
ypred.Pmcs<-c(table.S.1$n*prob.Pmcs)
ypred.Cmcs<-c(table.S.1$n*prob.Cmcs)
D.Lmcs<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Lmcs)+(table.S.1$n-table.S.1$y)*
      log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Lmcs))))  #deviance of logit
D.Pmcs<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Pmcs)+(table.S.1$n-table.S.1$y)*
      log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Pmcs))))  #deviance of probit
D.Cmcs<- c(2*sum(table.S.1$y*log(table.S.1$y/ypred.Cmcs)+(table.S.1$n-table.S.1$y)*
      log((table.S.1$n-table.S.1$y)/(table.S.1$n-ypred.Cmcs))))  #deviance of cloglog
P.D.Lmcs<-1-pchisq(D.Lmcs,df=8)
P.D.Pmcs<-1-pchisq(D.Pmcs,df=8)
P.D.Cmcs<-1-pchisq(D.Cmcs,df=8)
plot(log.load, table.S.1$y/table.S.1$n, pch=16, xlab="log.load of MCS",
      ylab="Proportion of fasteners that fail ", bty="L", axes=F)
axis(1, at=seq(7,9, .1))
axis(2, at=seq(0, 1, .1))
lines(log.load, prob.Lmcs,lty = 4,col=2)
lines(log.load, prob.Pmcs,lty = 5,col=3)
lines(log.load, prob.Cmcs,lty = 6,col=4)
legend(x=8.25, y=.35, legend=c("Logit","Probit","Clog-log"),lty=c(4,5,6),col=c(2,3,4), cex=.85,
      text.width=1, adj =c(0,.5))
```

#cor plot#

```
plot(log.load, table.S.1$y/table.S.1$n, pch=16, xlab="Log.LOAD",
        ylab="Proportion of fasteners that fail ", bty="L", axes=F)
 axis(1, at=seq(7,9, .1))
 axis(2, at=seq(0, 1, .1))
points (log.load, prob.Lwls,pch = 1,col=2)
points (log.load, prob.Pwls, pch = 2,col=3)
points (log.load, prob.Cwls, pch = 3,col=4)
points (log.load, fitted(fit.logit),pch = 4,col=2)
points (log.load, fitted(fit.probit),pch = 5,col=3)
points (log.load, fitted(fit.cloglog),pch = 6,col=4)
points (log.load, prob.Lmcs,pch= 7,col=2)
points (log.load, prob.Pmcs,pch = 8,col=3)
points (log.load, prob.Cmcs,pch = 9,col=4)
legend(x=8.3,y=.45,legend=c("Lwls","Pwls","Cwls","Lmle","Pmle","Cmle","Lmcs","Pmcs","Cm
cs"),pch=c(1,2,3,4,5,6,7,8,9),col=c(2,3,4), cex=.8, text.width=1, adj =c(0,.5))

data.frame(b.Lwls, b.Pwls, b.Cwls, summary(fit.logit)$coefficients,
        summary(fit.probit)$coefficients,summary(fit.cloglog)$coefficients,b.Lmcs,b.Pmcs,
        b.Cmcs)
data.frame(p, prob.Lwls, prob.Pwls, prob.Cwls,fitted(fit.logit),fitted(fit.probit),
        fitted(fit.cloglog),prob.Lmcs,prob.Pmcs,prob.Cmcs)
data.frame(log.load,table.S.1 , p, ypred.Lwls, fitted.Lmle=round(table.S.1$n*fitted(fit.logit),1),
        ypred.Lmcs, ypred.Pwls,fitted.Pmle = round(table.S.1$n*fitted(fit.probit),1), ypred.Pmcs,
        ypred.Cwls ,fitted.Cmle =round(table.S.1$n*fitted(fit.cloglog),1),ypred.Cmcs)
data.frame(D.Lwls, D.Pwls,D.Cwls,fit.logit$deviance, fit.probit$deviance,
        fit.cloglog$deviance,D.Lmcs, D.Pmcs, D.Cmcs )
data.frame(P.D.Lwls,P.D.Pwls,P.D.Cwls,P.D.Lmle, P.D.Pmle, P.D.Cmle, P.D.Lmcs,
        P.D.Pmcs, P.D.Cmcs)
```

ภาคผนวก ข

ข้อมูลที่ใช้ในงานวิจัย

ข้อมูลชุดที่ 1 จำนวนผู้ได้รับทดสอบเซรุ่มที่ให้ผลบวก

กลุ่มอายุ	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง
0-11 เดือน	0.5	10	3
1-2 y	1.5	10	1
2-4 y	3	29	5
5-9 y	7	69	39
10-14 y	12	51	31
15-19 y	17	15	8
≥ 20	30	108	91

ที่มา : Draper,Voller and Carpenter(1972)

ข้อมูลชุดที่ 1 เป็นข้อมูลผู้อาศัยในหมู่บ้าน Amozonas ประเทศบราซิล ปี 1971 ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ เพื่อตรวจสอบในช่วงเวลาในการฉีดเซรุ่มป้องกันมาลาเรีย ว่าเซรุ่มที่ให้ผลบวก หรือไม่ให้ผลบวก ข้อมูลจาก Draper,Voller and Carpenter(1972) ซึ่ง Draper ได้ใช้ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ในการวิเคราะห์ข้อมูล

ข้อมูลชุดที่ 2 ข้อมูลผู้สูบบุหรี่ที่มีอาการหอบ

กลุ่มอายุ	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง
20-24y	22	1952	104
25-29y	27	1791	128
30-34y	32	2113	231
35-39y	37	2783	378
40-44y	42	2274	442
45-49y	47	2393	593
50-54y	52	2090	649
55-59y	57	1750	631
60-64y	62	1136	504

ที่มา : Ashford and Sowden(1970)

ข้อมูลชุดที่ 2 เป็นข้อมูลของผู้สูบบุหรี่ที่ปราศจากสารกัมมันตภาพรังสีที่มีอายุระหว่าง 20-64 ปี ของบริษัท Coalminers ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของอายุ ว่ามีอาการหอบหรือไม่มีอาการหอบ เมื่อทำงานใน Coalminers ข้อมูลจาก Ashford and Sowden(1970)

ข้อมูลชุดที่ 3 ข้อมูลของเพศชายที่เป็นโรคหัวใจ

ความดันโลหิต	ค่ากลาง	ขนาดตัวอย่าง	ค่าตอบสนอง
<117	111.5	156	3
117-126	121.5	252	17
127-136	131.5	284	12
137-146	141.5	271	16
147-156	151.5	139	12
157-166	161.5	85	8
167-186	176.5	99	16
>186	191.5	43	8

ข้อมูลชุดที่ 3 เป็นข้อมูลผู้อาศัยเพศชาย อายุ 40-59 ปี ในเมือง 2 เมือง ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความดันโลจิต เพื่อตรวจสอบในช่วง 6 ปี ต่อเนื่องกันว่าเป็นโรคหัวใจ หรือไม่เป็นโรคหัวใจ ข้อมูลจาก Cornfield (1962) และ Agresti (1990) ซึ่งได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล

ที่มา : Cornfield (1962)

ข้อมูลชุดที่ 4 จำนวนการตายของแมลงเมื่อได้รับสารพิษแต่ละระดับ

log-dose	ขนาดตัวอย่าง	ค่าตอบสนอง
0.41	50	6
0.58	48	16
0.71	46	24
0.89	49	42
1.01	50	44

. ข้อมูลชุดที่ 4 เป็นข้อมูลปริมาณความเข้มข้นของสารพิษที่มีผลต่อการตายของแมลง โดยทำการเทค log ปริมาณสารพิษ ข้อมูลชุดนี้ทำการวิเคราะห์ด้วยตัวแบบโพรบิท ข้อมูลจาก Martin(1942) และ Finney(1971)

ที่มา : Martin(1942)

ข้อมูลชุดที่ 5 จำนวนการตายของแมลงเมื่อได้รับสารพิษแต่ละระดับ

CONC	ขนาดตัวอย่าง	ค่าตอบสนอง
0.0018	10	1
0.0022	10	3
0.0026	10	5
0.003	10	7
0.0034	10	8

ข้อมูลชุดที่ 5 เป็นข้อมูลทางชีววิทยาเกี่ยวกับการตายของแมลงเมื่อได้รับระดับความเข้มข้นที่ต่างกัน เป็นข้อมูลจริงของ Muhammad(1990) ซึ่ง Muhammad ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล และยังประมาณค่าพารามิเตอร์ด้วยวิธี WLE กล่าวคือ ควรใช้วิธี WLS กับข้อมูลจริงทางชีววิทยา

ที่มา : Muhammad(1990)

ข้อมูลชุดที่ 6 จำนวนการตายของแมลงเมื่อได้รับสารพิษแต่ละระดับ

log-dose	ขนาดตัวอย่าง	ค่าตอบสนอง
0.72	58	3
0.8	61	19
0.87	63	16
0.93	59	37
0.98	57	49
1.02	55	54
1.07	57	55
1.1	61	60

ข้อมูลชุดที่ 6 เป็นข้อมูลความเข้มข้นของแอมโมเนีย ที่มีผลต่อการตายของแมลง โดยทำการเทค log ปริมาณสารพิษเพื่อปรับค่าให้ตัวแปรอธิบายมีการแจกแจงแบบปกติ ข้อมูลจาก Strand(1930)

ที่มา : Strand(1930)

ข้อมูลชุดที่ 7 จำนวนหมุดที่เกิดความเสียหาย

Pressure Load	log-load	ขนาดตัวอย่าง	ค่าตอบสนอง
2500	7.824046	50	10
2700	7.901007	70	17
2900	7.972466	100	30
3100	8.039157	60	21
3300	8.101678	40	18
3500	8.160518	85	43
3700	8.216088	90	54
3900	8.268732	50	33
4100	8.318742	80	60
4300	8.36637	65	51

. ข้อมูลชุดที่ 7 เป็นข้อมูลการเสียหายของหมุดเจาะบนเครื่องบิน เมื่อระดับความกดอากาศเพิ่มขึ้นที่ละ 200 psi จาก 2500-4300 psi ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องของความกดอากาศ ว่าความกดอากาศที่ระดับต่างจะมีผลต่อการเสียหายหรือไม่เสียหายของหมุดเจาะบนเครื่องบิน ข้อมูลจาก Montgomery and Peck(1982) ได้ใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล ทางผู้วิจัยได้ทำการเทค log เพื่อปรับค่าตัวแปรอธิบายให้มีการแจกแจงแบบปกติ ข้อมูลที่ใช้ในงานวิจัยเป็นดังนี้

ที่มา : Montgomery and Peck(1982)

ข้อมูลชุดที่ 8 ข้อมูลการเลือก Reagan เป็นประธานาธิบดี

Political views	ขนาดตัวอย่าง	ค่าตอบสนอง
1	13	1
2	70	13
3	115	44
4	261	155
5	153	92
6	141	100
7	26	18

ข้อมูลชุดที่ 8 เป็นข้อมูลการสำรวจทางสังคม ปี 1982 ของคนผิวขาว ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่อง Political Views เพื่อดูว่าคนผิวขาวจะเลือก Reagan เป็นประธานาธิบดี เป็นตัวอย่างในแบบฝึกหัดโดยกำหนดให้เลือกใช้ตัวแบบโลจิทในการวิเคราะห์ข้อมูล

ที่มา : Clogg and Shockey (1988)

ข้อมูลชุดที่ 9 การสำรวจความคิดเห็นบทบาทของผู้หญิงที่มีต่อสังคม

Year of Edication	ขนาดตัวอย่าง	ค่าตอบสนอง
3	16	12
4	20	15
5	41	27
6	56	42
7	84	53
8	251	166
9	123	59
10	199	87
11	207	86
12	953	305
13	210	48
14	206	46
15	73	16
16	253	28
17	63	6
18	50	1

ที่มา : Haberman(1978)

ข้อมูลชุดที่ 9 เป็นข้อมูลการสำรวจความคิดเห็นเกี่ยวกับบทบาทของผู้หญิงที่มีต่อสังคม ทำการสำรวจทั้ง เพศหญิงและเพศชาย ถูกจัดกลุ่มเป็นหลายกลุ่มในเรื่องปีของการสำเร็จการศึกษา ต่อความคิดเห็นของการเห็นด้วยหรือไม่เห็นด้วยของ คำกล่าวที่ว่า "ผู้หญิงมีหน้าที่ดูแลบ้านและอนุญาตให้ทำงานนอกบ้านได้เหมือนผู้ชาย" ข้อมูลจาก Haberman(1978) ได้ทำการตัดข้อมูลปีสำเร็จการศึกษาตั้งแต่ 0-2 เนื่องจากcell นี้มีค่าเป็นศูนย์ และ 19-20 ออกเพราะเนื่องจากความถี่ใน cell นี้มีค่ากระโดดไม่คงที่

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวกุลพัชร หมื่นมา เกิดเมื่อวันที่ 1 พฤศจิกายน 2525 ที่ จ. เพชรบูรณ์ สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาสถิติ ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร เมื่อปีการศึกษา 2549 และเข้าศึกษาต่อในหลักสูตรสถิติศาสตรมหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2549