



เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิด
ในตัวแทนเชิงเส้นน้อยทั่วไป

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

โดย

นางสาวปานจิต วัฒนสารัช

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาสถิติประยุกต์

ภาควิชาสถิติ

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2552

ลิขสิทธิ์ของบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิด
ในตัวอย่างเชิงเส้นนัยทั่วไป

โดย

นางสาวปานจิต วัฒนสารัช

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาสถิติประยุกต์

ภาควิชาสถิติ

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2552

ลิขสิทธิ์ของบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

**CALIBRATION TECHNIQUES FOR CORRECTIONS OF MEASUREMENT ERROR AND
MISCLASSIFICATION ERROR IN GENERALIZED LINEAR MODELS**

By

Parnchit Wattanasaruch

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Statistics

Graduate School

SILPAKORN UNIVERSITY

2009

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์เรื่อง “เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิดในตัวเองเชิงเส้นนัยทั่วไป” เสนอ โดย นางสาวปานจิต วัฒนสารัช เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์

.....
(รองศาสตราจารย์ ดร.ศิริชัย ชินะตั้งกูร)
คณบดีบัณฑิตวิทยาลัย
วันที่.....เดือน..... พ.ศ.....

อาจารย์ที่ปรึกษาวิทยานิพนธ์
รองศาสตราจารย์วีรานันท์ พงศาภักดี

คณะกรรมการตรวจสอบวิทยานิพนธ์
มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

..... ประธานกรรมการ
(อาจารย์ ดร. ไพโรจน์ ขาวสิทธิวงษ์)
...../...../.....

..... กรรมการ
(รองศาสตราจารย์ ดร. กมล นุษบา)
...../...../.....

..... กรรมการ
(รองศาสตราจารย์วีรานันท์ พงศาภักดี)
...../...../.....

50304203 : สาขาวิชาสถิติประยุกต์

คำสำคัญ : การสอบเทียบการถดถอย/ ตัวแบบลوجิต/ ตัวแบบโพรบิต/ ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก

ปานจิต วัฒนสารัช : เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิดในตัวแบบเชิงเส้นนัยทั่วไป. อาจารย์ที่ปรึกษาวิทยานิพนธ์ : รศ.วีรานันท์ พงสาภักดี. 69 หน้า.

การศึกษาทางคลินิกวิทยาและระบาดวิทยาย่อยครั้งที่นำวิธีการสถิติ เรื่องการถดถอยเชิงเส้นและการถดถอยลوجิสติกมาใช้ ซึ่งอาจถูกกระทบโดยตัวแปรอธิบายตัวหนึ่งหรือมากกว่าที่มีความผิดพลาดจากการวัดค่า ทำให้ค่าประมาณที่ได้เกิดความเอนเอียงและส่งผลต่อการทดสอบสมมติฐานของพารามิเตอร์ วิธีสำหรับการปรับแก้ความผิดพลาดที่พบคือวิธีการสอบเทียบการถดถอย (Regression Calibration) คือการพยากรณ์หรือประมาณค่าจริงของตัวแปรอธิบายแบบต่อเนื่องจากค่าสังเกตภายใต้การวิเคราะห์การถดถอย งานวิจัยนี้เสนอเทคนิคการสอบเทียบใหม่ 4 วิธีสำหรับตัวแปรอธิบายจำแนกประเภทที่มีการจำแนกกลุ่มผิด โดยพัฒนาจากแนวคิดของวิธีการสอบเทียบการถดถอยนำไปสู่เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิดได้แก่ เทคนิคการสอบเทียบโพรบิต เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก เทคนิคการสอบเทียบลوجิสติกและเทคนิคการสอบเทียบลوجิต สำหรับการพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย 2 กลุ่มและใช้ค่าพยากรณ์กลุ่มนี้ในการวิเคราะห์จริง ภายใต้ตัวแบบเชิงเส้นนัยทั่วไป 3 ตัวแบบคือตัวแบบลوجิต ตัวแบบโพรบิตและตัวแบบคอมพลีเมนทารี ล็อก-ล็อกและเปรียบเทียบผลลัพธ์ด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE)

ผลการวิจัยพบว่าเทคนิคการสอบเทียบโพรบิตมีความถูกต้องมากที่สุด รองลงมาคือเทคนิคการสอบเทียบลوجิสติก เทคนิคการสอบเทียบลوجิตและเทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก ตามลำดับในเกือบทุกกรณี โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและพบว่าโดยทั่วไปตัวแบบโพรบิตเป็นตัวแบบที่เหมาะสมที่สุด นอกจากนี้เทคนิคการสอบเทียบโดยใช้ตัวแปรอธิบายเชิงกลุ่มให้ค่าประมาณพารามิเตอร์ที่ส่งผลต่อตัวแปรตอบสนองดีขึ้นเมื่อใช้เทคนิคการสอบเทียบโพรบิต เพื่อประมาณกลุ่มจริงของตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่มีความผิดพลาดจากการจำแนกกลุ่มก่อนการวิเคราะห์

ภาควิชาสถิติ

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2552

ลายมือชื่อนักศึกษา.....

ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์

50304203 : MAJOR : APPLIED STATISTICS

KEY WORDS : REGRESSION CALIBRATION/ LOGIT MODEL/ PROBIT MODEL/
COMPLEMENTARY LOG-LOG MODEL

PARNCHIT WATTANASARUCH : CALIBRATION TECHNIQUES FOR
CORRECTIONS OF MEASUREMENT ERROR AND MISCLASSIFICATION ERROR IN
GENERALIZED LINEAR MODELS. THESIS ADVISOR : ASSOC.PROF.VEERANUN
PONGSAPUKDEE. 69 pp.

The analyses of clinical and epidemiologic studies are often based on some kind of regression analysis, mainly linear or logistic regression. These analyses are often affected by the fact that one or more of the predictors are measures with error. Errors in the predictor are known to bias the estimates and hypotheses. One of the methods frequently used for adjusting for measurement error is the method of regression calibration. The idea here is to predict the true value of error-prone predictor from the observed data and use this predicted value in the regression analysis. This research proposes the four new calibration techniques for the misclassification explanatory variable. The methods are developed from the idea of the regression calibration leading to the calibration techniques namely probit, complementary log- log, logistic and logit calibration to predict the true value of binary predictor and use this predicted value in the three generalized linear models including logit, probit and complementary log- log models. The proposed methods and the models methods are compared by using the mean square error (MSE)

The research results show that the best three calibration technique is probit, logistic, logit, respectively, and complementary log- log is the worst for almost all cases. In addition, the probit model is the appropriate model. Moreover, the calibration technique with dichotomous explanatory variable yield the parameter estimates that improve the effect to the response variable when using probit calibration to predict the true group of the misclassification dichotomous explanatory variable before analyzing.

มหาวิทยาลัยศิลปากร ส่วนลิขสิทธิ์

Department of Statistics Graduate School, Silpakorn University Academic Year 2009
Student's signature
Thesis Advisor's signature

กิตติกรรมประกาศ

การศึกษาและเรียบเรียงวิทยานิพนธ์ฉบับนี้ สามารถเสร็จสมบูรณ์เป็นรูปเล่มและเสร็จ
ลุล่วงไปได้ด้วยดี ผู้เขียนได้รับความอนุเคราะห์และกรุณาจากรองศาสตราจารย์วีรพันธ์ พงศาภักดี
อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำแนะนำและช่วยเหลือในการเรียบเรียงวิทยานิพนธ์และแก้ไข
ข้อบกพร่องต่าง ๆ ด้วยความเมตตา เอาใจใส่อย่างยิ่ง

ขอกราบขอบพระคุณ อาจารย์ ดร. ไพโรจน์ ขาวสิทธิวงษ์และรองศาสตราจารย์ ดร.
กมล บุญบา กรรมการตรวจสอบวิทยานิพนธ์ที่ให้ความกรุณาแนะนำเพื่อแก้ไขข้อบกพร่อง เพื่อให้
วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากรที่กรุณาให้ความอนุเคราะห์ สนับสนุน
และอนุมัติให้ดำเนินงานวิจัยในครั้งนี้

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา คณาจารย์ ที่ให้กำลังใจและให้การสนับสนุน
ทางการศึกษา ตลอดจนเพื่อน ๆ ที่ให้คำแนะนำต่าง ๆ ในช่วงเวลาของการเรียนและการเรียบเรียง
วิทยานิพนธ์ตลอดมา และเป็นส่วนสำคัญที่ทำให้การเรียบเรียงงานวิจัยฉบับนี้เป็นไปด้วยดี

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญตาราง	ฅ
สารบัญภาพ	ญ
บทที่	
1 บทนำ.....	1
ความเป็นมาและความสำคัญของปัญหา.....	1
วัตถุประสงค์ของการวิจัย.....	4
ขอบเขตของการวิจัย	4
นิยามศัพท์เฉพาะ	6
ประโยชน์ที่คาดว่าจะได้รับ.....	6
2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง.....	7
ทฤษฎีที่เกี่ยวข้อง.....	7
งานวิจัยที่เกี่ยวข้อง	21
3 วิธีดำเนินการวิจัย	24
วิธีการวิจัย	24
ข้อมูลที่ใช้ในการวิจัย.....	24
การประมวลผลของการจำลองแบบ.....	24
สถิติที่ใช้ในการวิจัย.....	25
วิธีการจำลองแบบข้อมูล	25
วิธีการวิเคราะห์ข้อมูล	28
การกำหนดความผิดพลาดในการวิจัย.....	28
4 ผลการวิจัย.....	30
เทคนิคการสอบเทียบ 4 วิธีเพื่อปรับแก้ความผิดพลาดจากการจำแนกกลุ่มใน ตัวแปรอธิบาย.....	31
ผลกระทบจากตัวแบบ GLMs 3 ตัวแบบเมื่อตัวแปรอธิบายมีความผิดพลาด จากการจำแนกกลุ่ม.....	41

บทที่	หน้า
5	
สรุป อภิปรายผลและข้อเสนอแนะ	47
สรุปและอภิปรายผลการวิจัย	47
ข้อเสนอแนะของการวิจัย	48
ข้อเสนอแนะในการวิจัยครั้งต่อไป	48
บรรณานุกรม	49
ภาคผนวก	51
ประวัติผู้วิจัย	69

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สารบัญญัตราง

ตารางที่	หน้า
1 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลวจิต ภายใต $\sigma_U^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	31
2 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลวจิต ภายใต $\sigma_U^2 = 1$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	32
3 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลวจิต ภายใต $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	33
4 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต $\sigma_U^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	34
5 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต $\sigma_U^2 = 1$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	35
6 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	36
7 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต $\sigma_U^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	37
8 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต $\sigma_U^2 = 1$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	38
9 ความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลีเมนทารี ล็อก-ล็อก ภายใต $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธี.....	40
10 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมภายใต $\sigma_U^2 = 0.75$	41
11 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมภายใต $\sigma_U^2 = 1$	42
12 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมภายใต $\sigma_U^2 = 3$	43

สารบัญภาพ

ภาพที่	หน้า
1	กราฟแสดงโค้งของ $P(x)$ สำหรับตัวแบบลอจิตและตัวแบบโพรบิต ที่มีรูปแบบสมมาตรรอบจุด 0.5 12
2	กราฟแสดงโค้งของ $P(x)$ สำหรับตัวแบบคอมพลิเมนทารี ล็อก-ล็อก เมื่อค่าของ $P(x)$ เพิ่มขึ้นจาก 0 ก่อนข้างซ้ายแต่เข้าใกล้ 1 อย่างรวดเร็วหรือลดจาก 1 อย่างรวดเร็ว แต่เข้าใกล้ 0 ก่อนข้างซ้าย 13
3	ขั้นตอนของเทคนิคการสอบเทียบโพรบิต..... 19
4	ขั้นตอนของเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก..... 20
5	ขั้นตอนของเทคนิคการสอบเทียบลอจิสติก..... 20
6	ขั้นตอนของเทคนิคการสอบเทียบลอจิต..... 21
7	ขั้นตอนการจำลองแบบข้อมูล 29
8	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกเทคนิคการสอบเทียบ 3 วิธีที่ปรับแก้ ด้วยตัวแปรอธิบาย W_c แบบต่อเนื่อง ภายใต้ตามความแปรปรวนของ ความผิดพลาด ขนาดตัวอย่างและตัวแบบ GLMS 44
9	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกเทคนิคการสอบเทียบ 3 วิธีที่ปรับแก้ ด้วยตัวแปรอธิบาย W_g แบบ 2 กลุ่ม ภายใต้ตามความแปรปรวนของ ความผิดพลาด ขนาดตัวอย่างและตัวแบบ GLMS 45

บทที่ 1

บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

การศึกษาและการวิจัยทางคลินิกวิทยาและระบาดวิทยา นิยมนำวิธีการทางสถิติทั้งพรรณนาและการทดสอบสมมติฐานเชิงสถิติ รวมถึงการนำตัวแบบเชิงสถิติมาใช้สำหรับการวิเคราะห์ข้อมูล เช่น การถดถอยเชิงเส้นหรือการถดถอยลอจิสติก การถดถอยเชิงเส้นเป็นเครื่องมือที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนอง Y กับตัวแปรอธิบายแบบต่อเนื่อง X ที่มีความสัมพันธ์เชิงเส้นกัน ส่วนการถดถอยลอจิสติกเป็นเครื่องมือที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนอง Y จำแนกประเภทกับตัวแปรอธิบาย X แบบต่อเนื่องหรือแบบจำแนกประเภท การวิเคราะห์เหล่านี้ได้ถูกกระทบโดยตรงจากความจริงว่ามีตัวแปรอธิบายตัวหนึ่งหรือมากกว่ามีความผิดพลาดจากการวัดค่า ความผิดพลาดเหล่านี้อาจเป็นตัวแปรสุ่มหรือเป็นระบบ (ไม่สุ่ม) ถ้าต้องการวิเคราะห์ตัวแบบที่เกี่ยวกับตัวแปรอธิบาย X ซึ่งไม่สามารถสังเกตหรือวัดค่าได้โดยตรง แต่สามารถสังเกตหรือวัดค่าของตัวแปร W ได้ โดยที่ตัวแปร W และตัวแปรอธิบาย X สัมพันธ์กันเช่น $W = X + U$ เมื่อ U คือความผิดพลาดที่เกิดขึ้นและเป็นอิสระกับ X ดังนั้นการวิเคราะห์ตัวแบบที่เกี่ยวกับตัวแปรอธิบาย X โดยใช้ตัวแปร W จะนำมาสู่วิธีสอบเทียบ (Calibration) เพื่อช่วยในการประเมินผลที่ไม่ต้องมองข้ามปัญหาของค่าประมาณที่เกิดความเอนเอียงและส่งผลต่อการประมาณและการทดสอบสมมติฐานของพารามิเตอร์ ซึ่งในทางปฏิบัติมีผลกระทบอย่างมากต่อการนำไปใช้ โดยเฉพาะทางคลินิกวิทยาและระบาดวิทยา ในปัจจุบันมีหลายวิธีที่ถูกเสนอเพื่อรับมือกับปัญหานี้ สำหรับการถดถอยเชิงเส้นแบบธรรมดา (Ordinary regression) ถูกเสนอในหนังสือของ Fuller (1987) ส่วนการถดถอยลอจิสติกและตัวแบบไม่เชิงเส้นถูกเสนอในหนังสือของ Carroll, Ruppert and Stefanski (1995) คำตอบทั่วไปของปัญหาการประมาณนี้ในสถานการณ์ส่วนใหญ่จะพิจารณาวิธีการประมาณแบบภาวะน่าจะเป็นสูงสุด แต่ยังมียานวิจัยจำนวนน้อยมากที่ใช้วิธีการประมาณแบบ

ภาวะน่าจะเป็นสูงสุดเพื่อแก้ปัญหาความผิดพลาดของค่าวัด หนึ่งในเหตุผลหลัก คือ ปัญหาการคำนวณที่เกิดขึ้นเนื่องจากการหาค่าตอบของระบบสมการการประมาณ (Estimating equations) เกี่ยวกับ logistic-normal integrals วิธีที่พบสำหรับการปรับแก้ความผิดพลาดของค่าวัด คือ วิธีการสอบเทียบการถดถอย (Regression Calibration) แนวคิดของวิธีนี้ คือ การพยากรณ์หรือการประมาณค่าจริงของตัวแปรอธิบาย X แบบต่อเนื่องจากค่าสังเกต และใช้ค่าประมาณนี้ในการวิเคราะห์การถดถอย นี้คือแนวคิดง่าย ๆ ที่พิสูจน์ว่าให้ผลลัพธ์ที่เหมาะสมและอาจมีความน่าเชื่อถือใกล้เคียงกับวิธีอื่นที่มีความซับซ้อนกว่า (Thoresen and Laake 2000)

งานวิจัยที่ศึกษาเกี่ยวกับลักษณะของวิธีการประมาณพารามิเตอร์ ในสถานการณ์ที่ตัวแปรอธิบายมีความผิดพลาดจากการวัดค่าหรือจำแนกประเภทผิด เช่น Reade-Christopher and Kupper (1991) ได้เสนอตัวแบบที่ใช้ในการศึกษาติดตามผล (Follow up study) ที่ตัวแปรอธิบายมีความผิดพลาดจากการจำแนกกลุ่มในการถดถอยลอจิสติกและการถดถอยปัวซองแต่ไม่ได้ทำวิธีการสอบเทียบ Schafer (1993) ได้เสนอการวิเคราะห์ภาวะน่าจะเป็นสำหรับตัวแบบการถดถอยโพรบิต เมื่อมีตัวแปรอธิบายแบบต่อเนื่องบางตัวมีความผิดพลาดจากการวัด (Measurement error) โดยใช้วิธี EM Algorithm เพื่อประมาณภาวะน่าจะเป็นสูงสุด Thoresen and Laake (2000) ได้เสนอผลจากการจำลองแบบข้อมูลเพื่อเปรียบเทียบวิธีการประมาณพารามิเตอร์ 4 วิธีที่ใช้ในการปรับแก้ความผิดพลาดจากการวัดค่าในการถดถอยลอจิสติกแบบทวิภาค ภายใต้ตัวแปรอธิบายแบบต่อเนื่องทุกกรณีและพบว่าวิธีการสอบเทียบการถดถอยให้ผลดีพอ ๆ กับวิธีการซับซ้อน Gustafson and Le (2002) ทำการเปรียบเทียบความเอนเอียงที่เกิดจากความผิดพลาดของค่าวัดในตัวแปรอธิบายแบบต่อเนื่องกับความเอนเอียงที่เกิดจากความผิดพลาดของการจำแนกกลุ่มของตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มสำหรับถดถอยเชิงเส้นและการถดถอยลอจิสติก พบว่าการแบ่งกลุ่มสามารถลดความเอนเอียงที่เกิดขึ้นได้ Buzas, Tosteson and Stefanski (2003) ได้รวบรวมวรรณกรรมที่เกี่ยวข้องกับปัญหาและการสร้างตัวแบบที่แสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอธิบายแบบต่อเนื่องแบบต่าง ๆ สำหรับการเกิดความผิดพลาดจากการวัด Weller et al.(2007) ได้เสนอวิธีที่ใช้จัดการกับปัญหาความเอนเอียงที่เกิดขึ้นในค่าประมาณที่เกิดจากความผิดพลาดจากการวัดค่าในตัวแปรอธิบายแบบต่อเนื่อง ภายใต้ข้อสมมติของการมีตัวแปรแทน (Surrogates) มากกว่า 1 ตัวที่ใช้นิยามตัวแปรอธิบายแบบต่อเนื่องและใช้วิธีการสอบเทียบการถดถอยเพื่อปรับค่าประมาณของตัว

แปรอธิบายแบบต่อเนื่องมีความผิดพลาดจากการวัดค่า Lyles and Lin (2009) ได้เสนอวิธีการวิเคราะห์ความไว (Sensitivity) สำหรับตัวแปรอธิบายที่มีความผิดพลาดจากการจำแนกกลุ่มและตัวแปรอธิบายที่ไม่มีความผิดพลาดทั้งแบบต่อเนื่องและแบบทวิภาค ภายใต้ตัวแบบการถดถอยลอจิสติก จากงานวิจัยที่ผ่านมาพบว่าใช้วิธีการปรับแก้ความผิดพลาดด้วยวิธีการสอบเทียบการถดถอยเฉพาะตัวแปรอธิบายแบบต่อเนื่อง ซึ่งเป็นเทคนิคที่ไม่ซับซ้อนแต่ยังไม่พบงานวิจัยเกี่ยวกับการใช้เทคนิคการสอบเทียบสำหรับตัวแปรอธิบายแบบไม่ต่อเนื่องหรือแบบจำแนกประเภท มาแก้ปัญหาสำหรับตัวแปรอธิบายที่มีความผิดพลาดจากการจำแนกกลุ่มสำหรับตัวแบบเชิงเส้นน้อยทั่วไป (Nelder and Wedderburn 1972)

ดังนั้นงานวิจัยนี้จึงขยายขอบเขตการศึกษาจากงานวิจัยข้างต้น โดยเน้นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตอบสนองจำแนกประเภท (Y) ทวิภาคกับตัวแปรจำแนกประเภท (W_g) 2 กลุ่มภายใต้ตัวแบบ GLMs จำนวน 3 ตัวแบบที่ใช้ฟังก์ชันเชื่อมโยง (Link functions) แตกต่างกันและตัวแปร (W_g) 2 กลุ่มที่เกิดจากการนำค่าของตัวแปร W_c ที่มีความผิดพลาดจากการวัดค่ามาแบ่งกลุ่ม เนื่องจากตัวแปร W_c มีความคลาดเคลื่อนเช่นนี้ตัวแปร W_g จึงมีความผิดพลาดจากการจำแนกกลุ่มด้วย การวิเคราะห์ความสัมพันธ์จึงอาศัยหลายตัวแบบได้แก่ ตัวแบบแบบลอจิต (Logit link) และตัวแบบที่มีเทอมหลักคล้ายตัวแบบลอจิตคือตัวแบบโพรบิต (Probit model) และตัวแบบคอมพลีเมนทารี ล็อก-ล็อก (Complementary log-log) ส่วนตัวแปรอธิบายซึ่งมีความผิดพลาดจากการจำแนกกลุ่มนี้อาจส่งผลต่อค่าประมาณพารามิเตอร์และการทดสอบสมมติฐานตลอดจนต่อตัวแปรตอบสนองด้วย ผู้วิจัยจึงเสนอวิธีการปรับแก้ความผิดพลาดที่เกิดขึ้นนี้ก่อนการวิเคราะห์จริง โดยใช้เทคนิคการสอบเทียบ (Calibration techniques) 4 วิธีและพบว่าเทคนิคการสอบเทียบทั้ง 4 วิธีทำให้ค่าประมาณพารามิเตอร์ในตัวแบบดีขึ้นกว่าการไม่ใช้เทคนิคการสอบเทียบ เพื่อพยากรณ์หรือประมาณกลุ่มจริงก่อนการวิเคราะห์ งานวิจัยนี้เริ่มต้นพิจารณาเฉพาะผลกระทบต่อค่าประมาณพารามิเตอร์ (β_1) ที่แสดงอิทธิพลของตัวแปรอธิบายเพียง 1 ตัวเท่านั้น เนื่องจากการศึกษาและการวิจัยทางคลินิกวิทยาและระบาดวิทยามุ่งเน้นที่อิทธิพลของตัวแปรอธิบายที่ส่งผลต่อตัวแปรตอบสนอง หลังจากนั้นเปรียบเทียบเทคนิคการสอบเทียบและตัวแบบ GLMs ดังกล่าว เพื่อหาข้อสรุปจากวิธีการที่ง่าย ไม่ซับซ้อนและให้ผลดีสามารถนำไปใช้ได้จริง เทคนิคการสอบเทียบสร้างขึ้นมาเป็นการขยายแนวคิดของวิธีการสอบเทียบการถดถอยนำไปสู่

เทคนิคการสอบเทียบเพื่อใช้พยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย 2 กลุ่มและใช้ค่าพยากรณ์กลุ่มนี้ในการวิเคราะห์จริง เพื่อลดความเอนเอียงที่เกิดขึ้นในค่าประมาณพารามิเตอร์ ด้วยการพยากรณ์หรือการประมาณกลุ่มจริงจากตัวแปร W และตัวแปร W_g และเรียกเทคนิคการสอบเทียบ 4 วิธีที่เสนอนี้ว่าการสอบเทียบโพรบิต (Probit calibration) การสอบเทียบคอมพลิเมนต์ลอจิสติก (Complementary log-log calibration) การสอบเทียบลอจิสติก (Logistic calibration) และการสอบเทียบลอจิต (Logit calibration) ภายใต้ตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลิเมนต์ลอจิสติก ล็อก-ล็อก และเปรียบเทียบผลลัพธ์ของการประมาณค่าพารามิเตอร์ด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE)

2. วัตถุประสงค์ของการวิจัย

1. เสนอเทคนิคการสอบเทียบที่ไม่ซับซ้อน 4 วิธี สำหรับการปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบาย ด้วยตัวแปร W_c แบบต่อเนื่องและตัวแปร W_g เชิงกลุ่ม 2 กลุ่ม
2. เปรียบเทียบผลกระทบจากตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลิเมนต์ลอจิสติก ล็อก-ล็อก เมื่อตัวแปรอธิบายมีความผิดพลาดจากการจำแนกกลุ่ม

3. ขอบเขตของการวิจัย

3.1 ขอบเขตการจำลองแบบข้อมูล

1. กำหนดขนาดตัวอย่างเท่ากับ 100, 500 และ 1,000
2. กำหนดความแปรปรวนของความผิดพลาด (σ_U^2) เท่ากับ 0.75, 1 และ 3
3. กำหนดตัวแปรอธิบาย X_c แบบต่อเนื่องที่ไม่มีความผิดพลาดจากการวัด มีการแจกแจงแบบปกติ จำนวน 1 ตัวแปร ค่าเฉลี่ย 0 ความแปรปรวน 1
4. กำหนดตัวแปรอธิบายแบบ X_g แบบเชิงกลุ่ม 2 กลุ่มจากตัวแปรอธิบาย X แบบต่อเนื่อง โดยใช้เกณฑ์การแบ่งกลุ่ม คือ ค่าเฉลี่ยของ X_c เท่ากับ 0
5. กำหนดตัวแปรความผิดพลาดจากการวัด U มีค่าเฉลี่ย 0 ความแปรปรวน σ_U^2
6. กำหนดตัวแปร W_c แบบต่อเนื่องที่มีความผิดพลาดจากการวัด โดย $W_c = X_c + U$

7. กำหนดตัวแปร W_g แบบเชิงกลุ่ม 2 กลุ่ม จากตัวแปร W_c แบบต่อเนื่อง โดยใช้เกณฑ์การแบ่งกลุ่มจากค่าเฉลี่ย ของ W_c เท่ากับ 0

8. กำหนดพารามิเตอร์ของตัวแบบลอจิส ตัวแบบโพรบิตและตัวแบบคอมพลีเมนทารี ล็อก-ล็อก คือ $\beta_0 = -2.25, \beta_1 = 0.371$ (Thoresen and Laake 2000)

3.2 ขอบเขตตัวแบบที่ใช้ในการวิเคราะห์ข้อมูล

1. ตัวแบบลอจิส
2. ตัวแบบโพรบิต
3. ตัวแบบคอมพลีเมนทารี ล็อก-ล็อก

3.3 ขอบเขตความผิดพลาดของค่าวัด

1. กำหนดให้ X_c แทนตัวแปรอธิบายที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง
2. กำหนดให้ W_c แทนตัวแปรที่มีความผิดพลาดจากค่าวัด
3. กำหนดให้ลักษณะของตัวแบบความผิดพลาดของค่าวัดคือ Structural model โดยพิจารณา X_c เป็นตัวแปรสุ่มจากการแจกแจงแบบปกติ ค่าเฉลี่ย 0 ความแปรปรวน 1
4. กำหนดให้ความผิดพลาดที่เกิดขึ้นในตัวแปร W_c เป็นแบบ Nondifferential error หมายความว่าตัวแปร W_c ไม่มีสารสนเทศสำหรับการพยากรณ์ค่าของ Y นอกเหนือจากสารสนเทศที่อยู่ใน X_c และอาจเรียกตัวแปร W_c ในกรณีนี้ว่า Surrogate ของตัวแปรอธิบาย X_c
5. กำหนดให้ตัวแบบความผิดพลาดคือ Classical error models มีรูปแบบคือ $W_c = X_c + U$ เมื่อ U คือความผิดพลาดที่เกิดขึ้นมีค่าเฉลี่ยเท่ากับ 0 และเป็นอิสระกับ X_c

3.4 ขอบเขตตัวสถิติที่ใช้ในการวิเคราะห์ข้อมูล

ตัวสถิติที่ใช้ในการวิเคราะห์ข้อมูลคือความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของค่าประมาณพารามิเตอร์จากการวิเคราะห์ข้อมูลจำลอง M ครั้ง สูตรการคำนวณ MSE มีดังนี้

$$MSE = \frac{\sum_{m=1}^M (\hat{\theta}_m - \theta)^2}{M}$$

เมื่อ θ แทนพารามิเตอร์ความชันของตัวแบบเชิงเส้นน้อยที่สุด และ $M=1,000$

4. นิยามศัพท์เฉพาะ

1. ตัวแปร X^*_{wc} แทนตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่มาจาก การสอบเทียบด้วยตัวแปร W_c แบบต่อเนื่อง

2. ตัวแปร X^*_{wg} แทนตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่มาจาก การสอบเทียบด้วยตัวแปร W_g เชิงกลุ่ม 2 กลุ่ม

3. ตัวแปร X_c แทนตัวแปรอธิบายแบบต่อเนื่องที่ไม่มีความผิดพลาดจากการวัดค่า

4. ตัวแปร X_g แทนตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่ไม่มีความผิดพลาดจากการจำแนกกลุ่ม

5. ตัวแปร W_c แทนตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่า และ

$$W_c = X_c + U$$

6. ตัวแปร W_g แทนตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่มาจากตัวแปร W_c แบบต่อเนื่อง

7. ตัวแบบ GLMs แทนตัวแบบลอจิสต์ ตัวแบบ โพรบิตและตัวแบบคอมพลิเมนต์ารี ล็อก- ล็อก ที่มีตัวแปรอธิบายจำแนกประเภท 2 กลุ่ม

5. ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีการปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบาย สำหรับตัวแบบ GLMs

2. ได้ตัวแบบที่เหมาะสมสำหรับการวิเคราะห์ข้อมูลเชิงกลุ่ม 2 กลุ่มเมื่อตัวแปรอธิบายมีความผิดพลาดจากการจำแนกกลุ่ม

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. ทฤษฎีที่เกี่ยวข้อง

1.1 ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models : GLMs)

ตัวแบบจำนวนมากและส่วนใหญ่ที่ใช้ในการวิเคราะห์ข้อมูลจำแนกประเภทเป็นเพียงตัวแบบที่เป็นกรณีพิเศษของตัวแบบเชิงเส้นนัยทั่วไปซึ่งโดยหลักการในหลายกรณีเสนอโดย Nelder and Wedderburn (1972) และโดย McCullagh and Nelder (1989)

ตัวแบบเชิงเส้นนัยทั่วไปประกอบด้วยส่วนประกอบ 3 ส่วนคือ ส่วนประกอบเชิงสุ่ม (Random component) ซึ่งแสดงการแจกแจงความน่าจะเป็นของตัวแปรตอบสนอง ส่วนประกอบแบบมีระบบ (Systematic component) ซึ่งแสดงฟังก์ชันเชิงเส้นของตัวแปรอธิบายที่ใช้เป็นตัวพยากรณ์ และส่วนประกอบ Link function ซึ่งใช้อธิบายฟังก์ชันความสัมพันธ์ระหว่างส่วนประกอบแบบมีระบบกับค่าเฉลี่ยของส่วนประกอบเชิงสุ่มข้างต้น

1.2 ส่วนประกอบของตัวแบบเชิงเส้นนัยทั่วไป

ตัวแบบเชิงเส้นนัยทั่วไปเป็นตัวแบบที่ขยายจากตัวแบบเชิงเส้นแบบคลาสสิก (Classical linear model) โดยส่วนประกอบแรกที่เกี่ยวข้องกับตัวแปรเชิงสุ่มนั้น นอกจากมีการแจกแจงแบบปกติแล้วยังสามารถขยายไปสู่การแจกแจงในกลุ่มวงศเลขชี้กำลัง (Exponential family) ได้และส่วนประกอบของ Link function นอกจากจะใช้ Identical link แล้ว ก็ยังสามารถขยายให้ใช้กับฟังก์ชันเชื่อมโยงอื่น ๆ อีกหลายแบบที่เป็นฟังก์ชันแบบ Monotonic differentiable function ใด ๆ ก็ได้ ส่วนประกอบทั้ง 3 ส่วนมีรายละเอียดของแต่ละส่วนประกอบดังนี้

ส่วนประกอบที่ 1 ของ GLMs คือส่วนประกอบที่เกี่ยวข้องกับสมมติฐานของการแจกแจงของตัวแปรเชิงสุ่ม Y ที่เป็นตัวแปรตอบสนอง สมมติว่าค่าสังเกตจาก Y มีขนาด n หน่วยที่เป็นอิสระต่อกัน นั่นคือ $Y = (y_1, \dots, y_n)$ แต่ละส่วนประกอบของ Y คือ y_i ; $i=1, 2, \dots, n$ มีการแจกแจง

ในกลุ่มวงส์เลขชี้กำลัง ซึ่งอยู่ในรูปแบบของ

$$f_y(y; \theta, \phi) = \exp\left[\{y\theta - b\theta\} / a(\phi) + c(y, \phi)\right] \dots\dots\dots(1)$$

โดยที่ $a(\cdot)$, $b(\cdot)$ และ $c(\cdot)$ แทนฟังก์ชันต่าง ๆ ถ้าทราบ ϕ แล้ว (1) คือตัวแบบหนึ่งในกลุ่มวงส์เลขชี้กำลังที่มีพารามิเตอร์ θ แต่ไม่ทราบค่า ϕ สำหรับพารามิเตอร์ θ เรียกว่า Natural parameter ส่วน ϕ มักเรียกว่า Dispersion parameter และฟังก์ชัน $a(\phi)$ มักจะมีรูปแบบเป็น $a(\phi) = \phi / w_i$ โดยที่ w_i แทนน้ำหนักที่ทราบค่า เช่นเมื่อ \bar{y}_i แทนค่าเฉลี่ยของ n_i หน่วยที่เป็นอิสระต่อกัน จะใช้โดยทั่วไปว่า $w_i = n_i$ และเพื่อให้เข้าใจตัวแบบ (1) ได้ชัดเจนจะยกตัวอย่าง การแจกแจงแบบปกติ ซึ่งสามารถเขียนให้อยู่ในรูปแบบของ (1) ได้ดังนี้

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

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

ทำนองเดียวกันกับการแจกแจงแบบปกติ การแจกแจงแบบปัวซอง (Poisson) การแจกแจงแบบทวินาม (Binomial) และการแจกแจงในกลุ่มวงส์เลขชี้กำลังอื่น ๆ ก็สามารถเขียนให้อยู่ในรูปแบบของ (1) ได้เช่นกันและในกรณีที่ ϕ เป็นค่าคงที่ที่ทราบค่า (1) จะอยู่ในรูปแบบของ (2)

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) \exp\{y_i Q(\theta)\} \dots\dots\dots(2)$$

โดยที่ $Q(\theta)$ ใน (2) คือ $\theta / a(\phi)$ ใน (1)

$a(\theta)$ ใน (2) คือ $\exp\{-b(\theta) / a(\phi)\}$ ใน (1)

และ $b(y)$ ใน (2) คือ $\exp\{c(y, \phi)\}$ ใน (1)

จะเห็นว่าตัวแบบ (1) มีรูปแบบทั่วไปที่สามารถนำไปใช้ประโยชน์กับการแจกแจงในหลายรูปแบบ โดยเฉพาะสำหรับกลุ่มพารามิเตอร์ 2 ตัว (Two-parameter families) เช่นการแจกแจงแบบปกติและการแจกแจงแบบแกมมา (Gamma) ซึ่ง ϕ จะเป็นพารามิเตอร์ของความคลาดเคลื่อน (Nuisance parameter) ส่วนการแจกแจงสำหรับกลุ่มที่มีพารามิเตอร์ตัวเดียว (One-parameter families) เช่นการแจกแจงแบบปัวซอง การแจกแจงแบบทวินาม ไม่จำเป็นต้องใช้เทอม ϕ

ส่วนประกอบที่ 2 ของ GLMs คือส่วนประกอบแบบมีระบบทำหน้าที่เชื่อมเวกเตอร์ η โดยที่ $\eta = (\eta_1, \dots, \eta_N)'$ กับเซตของตัวแปรอธิบาย ให้มีรูปแบบเชิงเส้นดังนี้

$$\eta = X\beta \quad ; \quad \eta_j = \sum_i \beta_i X_{ij}; i=1, \dots, p; j=1, \dots, N$$

โดยที่ X แทนเมทริกซ์ของตัวแปรอธิบายที่ประกอบด้วยค่าสังเกตขนาด N อาจเรียก X ว่า Design matrix ที่มีขนาด $(N \times P)$

$$\beta \text{ แทนเวกเตอร์ของพารามิเตอร์ } (\beta_1, \dots, \beta_p)'$$

$$\eta \text{ แทนตัวพยากรณ์เชิงเส้น (Linear Predictor)}$$

ส่วนประกอบที่ 3 ของ GLMs คือ Link functions ต่าง ๆ สำหรับเชื่อมส่วนประกอบเชิงสุ่มและส่วนประกอบแบบมีระบบเข้าด้วยกัน เช่น

$$\text{ให้ } \mu_j = E(Y_j); j=1, \dots, N$$

$$\therefore \mu_j \text{ จะเกี่ยวข้องกับ } \eta_j \text{ ในรูปฟังก์ชันของค่าเฉลี่ยคือ } \eta_j = g(\mu_j)$$

โดยที่ g แทนฟังก์ชันแบบ Monotonic differentiable function ดังนั้นตัวแบบที่ต้องการจะเชื่อมโยงระหว่างค่าเฉลี่ยของค่าสังเกตของ Y กับตัวแปรอธิบายคือ

$$g(\eta_j) = \sum_i \beta_i X_{ij}; i=1, \dots, p; j=1, \dots, N$$

โดยที่ p แทนจำนวนของตัวแปรอธิบาย

ถ้า $g(\mu) = \mu$ จะได้ว่า $\eta_j = \mu_j$ คือ Identity link หรือเรียกอีกอย่างหนึ่งได้ว่า Canonical link โดยมีการแปลงค่าเฉลี่ยให้อยู่ในเทอมของพารามิเตอร์นั้นคือ

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

และ

$$Q(\theta_j) = \sum_i \beta_i X_{ij}; i=1, \dots, p$$

สรุปว่าตัวแบบ GLMs เป็นตัวแบบเชิงเส้นสำหรับค่าเฉลี่ยที่แปลงแล้วของตัวแปร ซึ่งมีการแจกแจงอยู่ในกลุ่มวงศ์เลขชี้กำลัง เพื่อแสดงให้เห็นถึงส่วนประกอบทั้งสามของตัวแบบ GLMs ให้ชัดเจนขึ้น จะยกตัวอย่างของตัวแบบ GLMs ที่นิยมใช้สำหรับตัวแปรตอบสนองจำแนกประเภท (Categorical response variables) เช่นตัวแบบลอจิต (Logit model) ซึ่งเป็นสับเซตของตัวแบบล็อกลิเนียร์และของตัวแบบเชิงเส้นที่วางนัยทั่วไปด้วย มีหลักการคือ ตัวแปรตอบสนองจำแนก

ประเภทอาจจัดเป็น 2 กลุ่ม เช่นค่าสังเกตของตัวแปรอาจจะถูกแบ่งเป็นกลุ่ม สำเร็จ หรือ ไม่สำเร็จ และแทนด้วยตัวเลขที่เป็นไปได้คือ 1 และ 0 ดังนั้นการแจกแจงแบบเบร์นูลลี (Bernoulli distribution) ของตัวแปรเชิงสุ่มแบบทวิภาคหรือ Binary นี้ จะให้ความน่าจะเป็นของ $P(Y = 1) = P$ และ $P(Y = 0) = 1 - P$ โดยที่ $P = E(Y)$

เมื่อ Y_i มีการแจกแจงแบบเบร์นูลลีด้วยพารามิเตอร์ P ฟังก์ชันความหนาแน่นของความน่าจะเป็น คือ

$$\begin{aligned} f(y_i; P_i) &= P_i^{y_i} (1 - P_i)^{1 - y_i} \\ &= (1 - P_i) \left[\frac{P_i}{1 - P_i} \right]^{y_i} \\ &= (1 - P_i) \exp \left[y_i \log \left(\frac{P_i}{1 - P_i} \right) \right] \end{aligned}$$

โดยที่ $y_i = 0, 1$ การแจกแจงข้างต้นอยู่ในกลุ่มเอกซ์โปเนนเชียล โดยมีเทอมของ

$$Q(P) = \log [P / (1 - P)]$$

ซึ่งเป็น Log odds of response 1 หรือ เรียกว่า Logit of P ในตัวแบบที่ใช้ฟังก์ชันเชื่อมโยงแบบ Logit link นี้เรียกว่า Logit model

1.3 ตัวแบบลอจิต (Logit model)

ตัวแบบลอจิต (Logit model) เป็นวิธีการทางสถิติสำหรับการวิเคราะห์ข้อมูล เมื่อตัวแปรตอบสนอง Y แบบเชิงกลุ่ม 2 กลุ่มหรือมากกว่า 2 กลุ่ม ส่วนตัวแปรอธิบาย X เป็นแบบจำแนกประเภท แต่ถ้าตัวแปรอธิบาย X เป็นแบบต่อเนื่อง เราเรียกว่าตัวแบบลอจิสติก (วีรานันท์ 2544)

กำหนดให้ $P(x) = E(Y | x)$ แทนค่าเฉลี่ยหรือค่าคาดหวังแบบมีเงื่อนไขของ Y เมื่อกำหนดค่าของ x โดยที่ Y เป็นตัวแปรตอบสนองแบบ Binary หรือ Dichotomous ส่วน x เป็นตัวแปรอธิบายจำแนกประเภท รูปแบบพื้นฐานของตัวแบบลอจิต คือ

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

ซึ่งเป็นรูปแบบที่แสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอธิบาย 1 ตัว พิจารณา $Y_i \sim \text{bin}(n_i, P(x)); i = 1, \dots, N$ โดยที่ $P(x)$ แทนความน่าจะเป็นของ $Y = 1$ จุดประสงค์คือต้องการหาความสัมพันธ์เชิงเส้นระหว่าง $P(x)$ กับ x ซึ่ง $0 \leq P(x) \leq 1$ ส่วน $\beta_0 + \beta_1 x$ ไม่จำเป็นต้องอยู่ในช่วง $(0, 1)$ จึงต้องหาฟังก์ชันอื่นมาสร้างตัวแบบที่แสดงความสัมพันธ์ของ $P(x)$

กับ x ใหม่ ในทางปฏิบัติจะต้องแปลงเทอม $P(x)$ ข้างต้นโดยเรียกว่า การแปลงลอจิต (Logit transformation) ให้อยู่ในเทอมของ

$$Q(x) = \log \left[\frac{P(x)}{1-P(x)} \right] = \text{logit ของ } P(x)$$

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

$$\log \left[\frac{P(x)}{1-P(x)} \right] = \beta_0 + \beta_1 \text{ เมื่อ } X = 1 \text{ หรือ } 0 \dots\dots\dots(3)$$

ตัวแบบ (3) คือ Logarithm ของ Odds ในรูปแบบความสัมพันธ์เชิงเส้นกับ x เรียกตัวแบบ (3) นี้ว่า ตัวแบบลอจิต (Logit model) เมื่อ X เป็นตัวแปรจำแนกประเภท (ในที่นี้มี 2 กลุ่ม) ตัวแบบ (3) มีรูปแบบเช่นเดียวกับ Odds ของตัวแปรตบสองเมื่อ $Y=1$ ด้วย

$$P(x) = [1-P(x)]e^{\beta_0+\beta_1x}$$

$$e^{\beta_0+\beta_1x} = P(x)[1+e^{\beta_0+\beta_1x}]$$

นั่นคือ $P(x) = \frac{e^{\beta_0+\beta_1x}}{1+e^{\beta_0+\beta_1x}}$ หรือ $P(x) = \frac{\exp(\beta_0 + \beta_1x)}{1+\exp(\beta_0 + \beta_1x)}$

ประโยชน์ของตัวแบบลอจิต คือ การตีความหมายในเทอมของ Odds โดยพบว่าในกรณีที่มีตัวแปรอธิบาย x เพียง 1 ตัวแปร Odds จะเพิ่มขึ้นเป็น $\exp(\beta_1)$ ณ ทุก ๆ ค่า x แบบต่อเนื่องที่เพิ่มขึ้น 1 หน่วย แต่ถ้า X เป็นตัวแปรจำแนกประเภท การตีความหมายตัวแบบลอจิตจะตีความหมายในเทอมของอัตราส่วน Odds แทน Odds และ Odds ของ $Y=1$ คือตัวแบบ

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

ซึ่งเป็น Link function สำหรับตัวแบบลอจิตให้มีรูปแบบของตัวแบบเชิงเส้นวางนัยทั่วไป สำหรับในกรณีที่มีตัวแปร x หลายตัว การวิเคราะห์ต่าง ๆ และการตีความหมายในเทอมของ Odds สามารถทำได้ในทำนองเดียวกันกับที่กล่าวข้างต้น

1.4 ตัวแบบโพรบิต (Probit model)

ตัวแบบโพรบิต (Probit model) คือตัวแบบที่มีเทอมหลักคล้ายตัวแบบลอจิต ซึ่งเป็นตัวแบบที่มีฟังก์ชันความหนาแน่นสะสม (CDF) อยู่ในรูปแบบของ $P(x) = F(\beta_0 + \beta_1 X)$ หรือ $\Phi^{-1}[P(x)] = \beta_0 + \beta_1 X$ ถ้าให้ G แทน CDF ของการแจกแจงแบบปกติ จะพบว่า

$P(x) = G(x) = \Phi[(x - \mu) / \sigma]$ โดยที่ Φ แทน CDF แบบปกติมาตรฐาน ซึ่งสามารถแปลงให้อยู่ในเทอมของ $P(x) = F(\beta_0 + \beta_1 X)$ โดยที่ $F = \Phi$, $\beta_0 = -\mu / \sigma$ และ $\beta_1 = 1 / \sigma$ การแปลงตัวแบบข้างต้นจะมีผลให้ตัวแบบมีรูปแบบใหม่ซึ่งเรียกว่า ตัวแบบโพรบิต (Pobit model) ดังใน (4) (Agresti 2002 และ วีรานันท์ 2544)

$$\Phi^{-1}[P(x)] = \beta_0 + \beta_1 X \dots\dots\dots(4)$$

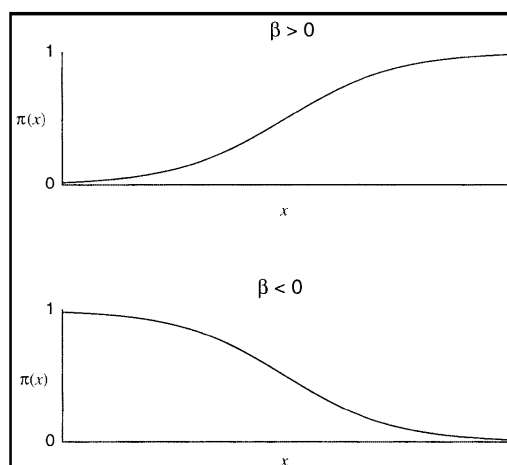
1.5 ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก (Complementary log-log model)

ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก (Complementary log-log model) เป็นส่วนขยายจากตัวแบบลอจิตและตัวแบบโพรบิต เมื่อค่าของ $P(x)$ เพิ่มจาก 0 ก่อนข้างเข้าแต่มีค่าเข้าใกล้ 1 อย่างรวดเร็ว สำหรับฟังก์ชัน link ที่ใช้สำหรับตัวแบบลอจิตและตัวแบบโพรบิตจะมีคุณสมบัติสมมาตร (symmetric) รอบค่า 0.5 หรือ $\text{link}[P(x)] = -\text{link}[1 - P(x)]$ นั่นคือ

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

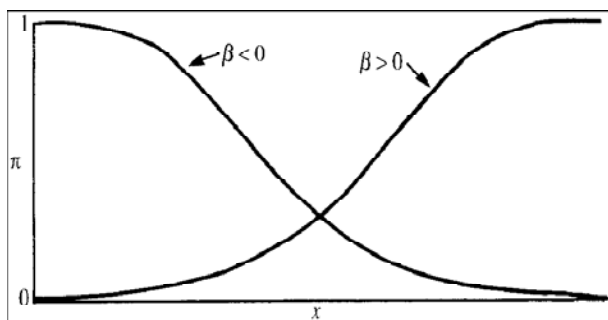
หมายความว่าโค้งของ $P(x)$ สำหรับตัวแบบลอจิตและตัวแบบโพรบิตมีรูปแบบสมมาตรรอบจุด 0.5

โดยเฉพาะ $P(x)$ จะมีค่าเข้าใกล้ 0 ด้วยอัตราที่เท่ากับเมื่อ $P(x)$ มีค่าเข้าใกล้ 1 ดังภาพที่ 1



ภาพที่ 1 กราฟแสดงโค้งของ $P(x)$ สำหรับตัวแบบลอจิตและตัวแบบโพรบิตที่มีรูปแบบสมมาตรรอบจุด 0.5

แต่ถ้าค่าของ $P(x)$ เพิ่มขึ้นจาก 0 ค่อนข้างช้าแต่เข้าใกล้ 1 อย่างรวดเร็วหรือลดจาก 1 อย่างรวดเร็วแต่เข้าใกล้ 0 ค่อนข้างช้า ตัวแบบลอจิสต์และตัวแบบโพรบิตจะไม่เหมาะสมกับข้อมูลควรใช้ตัวแบบอื่นคือ ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อกและลักษณะของกราฟ $P(x)$ สำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก ดังภาพที่ 2



ภาพที่ 2 กราฟแสดงโค้งของ $P(x)$ สำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก เมื่อค่าของ $P(x)$ เพิ่มขึ้นจาก 0 ค่อนข้างช้าแต่เข้าใกล้ 1 อย่างรวดเร็วหรือลดจาก 1 อย่างรวดเร็วแต่เข้าใกล้ 0 ค่อนข้างช้า

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

และลักษณะของกราฟของ $P(x)$ ควรใช้เส้นโค้งของฟังก์ชัน (5) คือ

$$P(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)] \dots\dots\dots (5)$$

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

ซึ่งมีรูปแบบไม่สมมาตรคือ $P(x)$ มีค่าลดจาก 1 รวดเร็วกว่าการเข้าใกล้ 0 โดยฟังก์ชัน (5) นำไปสู่ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก ใน (6) คือ (วีรพันธ์ 2544)

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

$$\log[-\log(1 - P(x))] = \beta_0 + \beta_1 x \dots\dots\dots (6)$$

1.6 ความผิดพลาดของค่าวัด (Measurement error)

ตัวแบบความผิดพลาดของค่าวัด พิจารณาภายใต้ตัวแบบสำหรับตัวแปรตอบสนอง Y ในเทอมของตัวแปรอธิบาย ซึ่งตัวแปรอธิบายที่ใช้ถูกแบ่งออกเป็น 2 ชนิดคือ Z แทนตัวแปรอธิบายที่ไม่มีความผิดพลาดของค่าวัดและ X แทนตัวแปรอธิบายที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง ลักษณะของปัญหาความผิดพลาดของค่าวัดคือ สามารถสังเกตตัวแปร W ซึ่งมี

ความสัมพันธ์กับ X ได้ และพารามิเตอร์ในตัวแบบที่แสดงความสัมพันธ์ระหว่าง Y กับ (Z, X) ไม่สามารถประมาณค่าได้โดยตรงจากการสร้างตัวแบบแสดงความสัมพันธ์ระหว่าง Y กับ (Z, X) วัตถุประสงค์ของการสร้างตัวแบบความผิดพลาดก็คือการหาค่าประมาณที่ไม่มีความเอนเอียงจากตัวแบบสำหรับ Y ในเทอมของ (Z, W) การใช้ W แทน X แต่ไม่มีการปรับแก้ความผิดพลาดที่เกิดขึ้นส่งผลให้ค่าประมาณที่ได้เกิดความเอนเอียงบางครั้งอาจส่งผลอย่างรุนแรง ในการประเมินความผิดพลาดของค่าวัดจำเป็นต้องทราบชนิดและลักษณะของความผิดพลาดรวมทั้งแหล่งที่มาของข้อมูลที่ใช้สำหรับความผิดพลาดแต่ละประเภทด้วย (Carroll, Ruppert and Stefanski 1995)

1.7 Functional and Structural Models

การกำหนดลักษณะของตัวแบบความผิดพลาดของค่าวัดตามคุณสมบัติของตัวแปรอธิบาย X แบ่งได้ 2 ประเภทคือ Classical functional models เมื่อ X คือค่าคงที่ที่ไม่ทราบค่าและ Classical structural models เมื่อ X คือตัวแปรสุ่ม (Carroll, Ruppert and Stefanski 1995)

1.8 Differential and Non-differential Error

ความผิดพลาดใน W ที่ใช้แทนค่าวัดของ X คือ Nondifferential error ถ้าการแจกแจงแบบมีเงื่อนไขของ Y เมื่อกำหนด (Z, X, W) คือการแจกแจงแบบมีเงื่อนไขของ Y เมื่อกำหนด (Z, X) นั่นคือ $f_{Y|zwx} = f_{Y|zx}$ หรือกล่าวอีกนัยหนึ่งว่า W ไม่มีสารสนเทศสำหรับการพยากรณ์ค่าของ Y นอกเหนือจากสารสนเทศที่อยู่ใน Z และ X ดังนั้น W คือ Surrogate และความผิดพลาดใน W ที่ใช้แทนค่าวัดของ X คือ Differential error ถ้า $f_{Y|zwx} \neq f_{Y|zx}$ (Carroll, Ruppert and Stefanski 1995)

1.9 Error Models

การกำหนดตัวแบบสำหรับความผิดพลาดแบ่งเป็น 2 ประเภทดังนี้ (Buzas, Tosteson and Stefanski 2003)

1. Classical error models คือตัวแบบทางสถิติพื้นฐานสำหรับกรณีที่ W คือค่าวัดของ X มีรูปแบบคือ $W = X + U$ เมื่อ U คือความผิดพลาดที่เกิดขึ้นมีค่าเฉลี่ยเท่ากับ 0 และเป็นอิสระกับ X ตัวแบบนี้เป็นรูปแบบที่ง่ายที่สุดและเหมาะสมที่สุดเมื่อต้องการทราบค่า X ที่แน่นอนแต่ไม่สามารถทำได้เนื่องจากความผันแปรต่าง ๆ ที่เกิดขึ้นในการวัดค่าซึ่งอาจมาจากเครื่องมือที่ใช้ในการวัดหรือปัจจัยอื่น ๆ

2. Berkson error models เป็นตัวแบบความผิดพลาดที่ถูกใช้เมื่อ X มีการเปลี่ยนแปลงรอบ ๆ ค่าของ W มีรูปแบบคือ $X = W + U$ เมื่อ U คือความผิดพลาดที่เกิดขึ้นมีค่าเฉลี่ยเท่ากับ 0 และเป็นอิสระกับ W

การกำหนด Error Models ที่ใช้ในการวิเคราะห์ข้อมูลขึ้นอยู่กับสถานการณ์และข้อมูลที่มี เช่น เมื่อต้องการทราบระดับความดันโลหิตระยะยาวของหน่วยตัวอย่าง ค่าที่วัดได้ ณ เวลาหนึ่งเป็นค่าวัดที่ถูกรบกวนด้วยความผิดพลาดเช่น เครื่องมือ การอ่านค่าของผู้บันทึก สภาวะร่างกายของผู้วัด ดังนั้นตัวแบบ Classical error models จึงมีความเหมาะสม ส่วน Berkson error models ถูกใช้กรณีที่มีการศึกษาในห้องปฏิบัติการ เช่น การอบวัตถุดิบในเตาอบ ณ อุณหภูมิที่กำหนด W ถูกกำหนดโดยอุปกรณ์ควบคุมอุณหภูมิให้คงที่ แม้ว่าจะกำหนดอุณหภูมิที่ W แต่อุณหภูมิจริงในเตาอบ X จะมีการเปลี่ยนแปลงไปตาม W เนื่องจากอุปกรณ์ควบคุมอุณหภูมิไม่ได้มีความสมบูรณ์ 100%

1.10 แหล่งข้อมูล

แหล่งข้อมูลสามารถแบ่งได้เป็น 2 กลุ่ม (Carroll, Ruppert and Stefanski 1995)

แหล่งข้อมูลภายใน (Internal data) หมายถึงข้อมูลที่ได้จากส่วนหนึ่งของข้อมูลที่

ศึกษาสามารถแบ่งเป็นกลุ่มย่อยได้ 3 ชนิดคือ

1. ข้อมูลจริง (Validation data) หมายถึงข้อมูลที่สามารถสังเกตหรือวัดค่า X ได้โดยตรง
2. ข้อมูลการวัดซ้ำ (Replication data) หมายถึงข้อมูลที่สามารถวัดซ้ำได้ของ W
3. ข้อมูลที่มีส่วนช่วย (Instrumental data) หมายถึงข้อมูลจากตัวแปรอื่น ๆ ที่

สามารถสังเกตได้นอกเหนือจาก W

แหล่งข้อมูลภายนอก (External data or Independent studies) หมายถึงข้อมูลที่ได้จากการศึกษาอื่น ๆ สามารถแบ่งเป็นกลุ่มย่อยได้ 3 ชนิดคือ

1. ข้อมูลจริง (Validation data) หมายถึงข้อมูลที่สามารถสังเกตหรือวัดค่า X ได้โดยตรง
2. ข้อมูลการวัดซ้ำ (Replication data) หมายถึงข้อมูลที่สามารถวัดซ้ำได้ของ W

3. ข้อมูลที่มีส่วนช่วย (Instrumental data) หมายถึงข้อมูลจากตัวแปรอื่น ๆ ที่สามารถสังเกตได้นอกเหนือจาก W

1.11 การสอบเทียบการถดถอย (Regression Calibration)

การสอบเทียบการถดถอยเป็นแนวคิดที่สามารถจัดการกับความเอนเอียงที่เกิดขึ้นในการประมาณพารามิเตอร์และเป็นวิธีที่นำมาใช้ในตัวแบบการถดถอย ตัวแบบเชิงเส้นและตัวแบบไม่เชิงเส้น ในกรณีที่ตัวแปรอธิบายเป็นแบบต่อเนื่อง ถูกศึกษาเป็นครั้งแรกในงานวิจัยเกี่ยวกับ Proportional hazards regression เมื่อ X เป็นตัวแปรแบบต่อเนื่องของ Prentice (1982) การสอบเทียบการถดถอยมีประโยชน์อย่างมากในการคำนวณค่าของ X แบบต่อเนื่อง แนวคิดคือการประมาณค่าของ X ที่สังเกตไม่ได้ด้วย X^* แบบต่อเนื่อง ซึ่ง X^* คือค่าพยากรณ์ของ X ที่ได้จากการถดถอยของ X บน (Z,W) การสร้างตัวแบบและการประมาณการถดถอยของ X บน (Z,W) ต้องการข้อมูลเพิ่มเติม ข้อมูลดังกล่าวอาจมาจากภายนอกหรือเป็นข้อมูลจากภายในหรืออาจเป็นข้อมูลที่มีการวัดซ้ำ พารามิเตอร์ที่ต้องการจะถูกประมาณโดยการถดถอยของ Y บน (Z, X^*) การสอบเทียบการถดถอยสามารถสรุปขั้นตอนของวิธีการสอบเทียบการถดถอยได้ 2 ขั้นตอนคือ การประมาณการถดถอยของ X บน (Z,W) เพื่อหาค่าของ X^* และการถดถอยของ Y บน (Z, X^*) เพื่อหาค่าประมาณพารามิเตอร์ (Buzas, Tosteson and Stefanski 2003)

1.12 เทคนิคการสอบเทียบโพรบิต (Probit Calibration)

เทคนิคการสอบเทียบโพรบิต เป็นเทคนิคการประมาณค่าตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม (x_g) ที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง ด้วยตัวแปรอธิบายแบบต่อเนื่อง (w_c) หรือแบบเชิงกลุ่ม 2 กลุ่ม (w_g) ที่สามารถสังเกตหรือวัดค่าได้ ภายใต้ตัวแบบโพรบิต ประกอบด้วยขั้นตอน ดังนี้

1. หาค่า $P(x_g = 1 | w_c)$ หรือ $P(x_g = 1 | w_g)$ ภายใต้ตัวแบบ

โพรบิต คือ

$$\Phi^{-1}[P(x_g = 1 | w_c)] = \beta_0 + \beta_1 w_c$$

หรือ

$$\Phi^{-1}[P(x_g = 1 | w_g)] = \beta_0 + \beta_1 w_g$$

โดยที่ x_g แทนตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรงอาจเป็นข้อมูลจริง (Validation data) ข้อมูลที่มีการวัดซ้ำ (Replication data) หรือข้อมูลที่มีส่วนช่วย (Instrumental data)

2. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ Cutpoint

3. สร้างตัวแปร X^*_{wc} โดยที่ $X^*_{wc} = 1$ ถ้า $P(x_g = 1 | w_c) > c$ และ $X^*_{wc} = 0$ กรณีอื่น ๆ หรือสร้างตัวแปร X^*_{wg} โดยที่ $X^*_{wg} = 1$ ถ้า $P(x_g = 1 | w_g) > c$ และ $X^*_{wg} = 0$ กรณีอื่น ๆ

ขั้นตอนของเทคนิคการสอบเทียบโพรบิต 1-3 แสดงไว้ในภาพที่ 3

1.13 เทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก (Complementary log-log Calibration)

เทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก เป็นเทคนิคการประมาณค่าตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม (x_g) ที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง ด้วยตัวแปรอธิบายแบบต่อเนื่อง (w_c) หรือแบบเชิงกลุ่ม 2 กลุ่ม (w_g) ที่สามารถสังเกตหรือวัดค่าได้ ภายใต้ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก ประกอบด้วยขั้นตอน ดังนี้

1. หาค่า $P(x_g = 1 | w_c)$ หรือ $P(x_g = 1 | w_g)$ ภายใต้ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก คือ

$$\log[-\log(1 - P(x_g = 1 | w_c))] = \beta_0 + \beta_1 w_c$$

หรือ

$$\log[-\log(1 - P(x_g = 1 | w_g))] = \beta_0 + \beta_1 w_g$$

โดยที่ x_g แทนตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรงอาจเป็นข้อมูลจริง (Validation data) ข้อมูลที่มีการวัดซ้ำ (Replication data) หรือข้อมูลที่มีส่วนช่วย (Instrumental data)

2. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ Cutpoint

3. สร้างตัวแปร X^*_{wc} โดยที่ $X^*_{wc} = 1$ ถ้า $P(x_g = 1 | w_c) > c$

และ $X^*_{wc} = 0$ กรณีอื่น ๆ และสร้างตัวแปร X^*_{wg} โดยที่ $X^*_{wg} = 1$ ถ้า

$P(x_g = 1 | w_g) > c$ และ $X^*_{wg} = 0$ กรณีอื่น ๆ

ขั้นตอนของเทคนิคการสอบเทียบคอมพลิเมนต์ารี ล็อก-ล็อก 1-3 แสดงไว้ในภาพที่ 4

1.14 เทคนิคการสอบเทียบลอจิสติก (Logistic Calibration)

เทคนิคการสอบเทียบลอจิสติก เป็นเทคนิคการประมาณค่าตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม (x_g) ที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง ด้วยตัวแปรอธิบายแบบต่อเนื่อง (w_c) หรือแบบเชิงกลุ่ม 2 กลุ่ม (w_g) ที่สามารถสังเกตหรือวัดค่าได้ ภายใต้ตัวแบบลอจิสติก ประกอบด้วยขั้นตอน ดังนี้

1. หาค่า $P(x_g = 1 | w_c)$ หรือ $P(x_g = 1 | w_g)$ ภายใต้ตัวแบบลอจิสติก คือ

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

โดยที่ x_g แทนตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรงอาจเป็นข้อมูลจริง (Validation data) ข้อมูลที่มีการวัดซ้ำ (Replication data) หรือข้อมูลที่มีส่วนช่วย (Instrumental data)

2. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ Cutpoint

3. สร้างตัวแปร X^*_{wc} โดยที่ $X^*_{wc} = 1$ ถ้า $P(x_g = 1 | w_c) > c$ และ $X^*_{wc} = 0$ กรณีอื่น ๆ

ขั้นตอนของเทคนิคการสอบเทียบลอจิสติก 1-3 แสดงไว้ในภาพที่ 5

1.15 เทคนิคการสอบเทียบลอจิต (Logit Calibration)

เทคนิคการสอบเทียบลอจิต เป็นเทคนิคการประมาณค่าตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม (x_g) ที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรง ด้วยตัวแปรอธิบายแบบต่อเนื่อง (w_c) หรือแบบเชิงกลุ่ม 2 กลุ่ม (w_g) ที่สามารถสังเกตหรือวัดค่าได้ ภายใต้ตัวแบบลอจิต ประกอบด้วยขั้นตอน ดังนี้

1. หาค่า $P(x_g = 1 | w_c)$ หรือ $P(x_g = 1 | w_g)$ ภายใต้ตัวแบบลอจิสต์ คือ

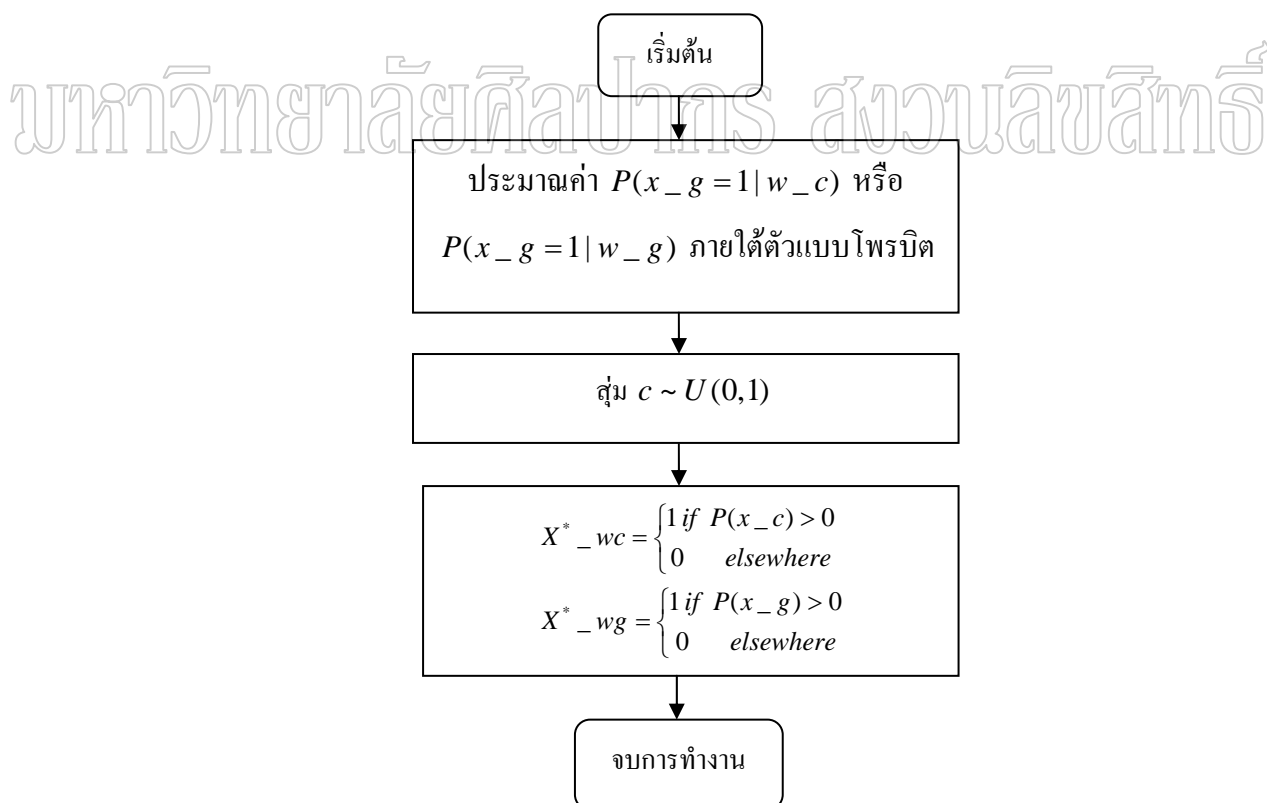
$$\log \left[\frac{P(x_g = 1 | w_c)}{1 - P(x_g = 1 | w_c)} \right] = \beta_0 + \beta_1 w_g$$

โดยที่ x_g แทนตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่ไม่สามารถสังเกตหรือวัดค่าได้โดยตรงอาจเป็นข้อมูลจริง (Validation data) ข้อมูลที่มีการวัดซ้ำ (Replication data) หรือข้อมูลที่มีส่วนช่วย (Instrumental data)

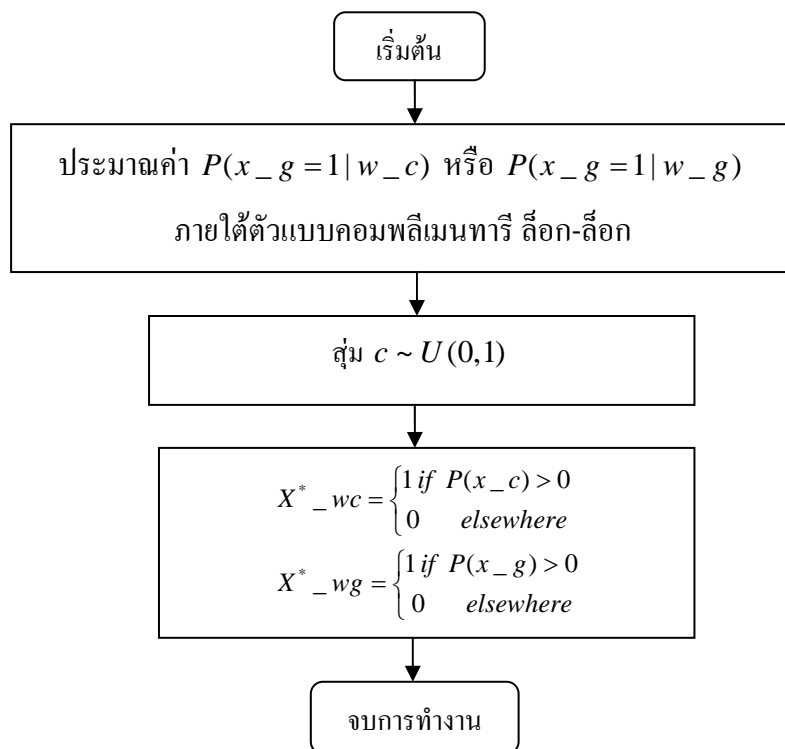
2. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ Cutpoint

3. สร้างตัวแปร X^*_{wg} โดยที่ $X^*_{wg} = 1$ ถ้า $P(x_g = 1 | w_g) > c$ และ $X^*_{wg} = 0$ กรณีอื่น ๆ

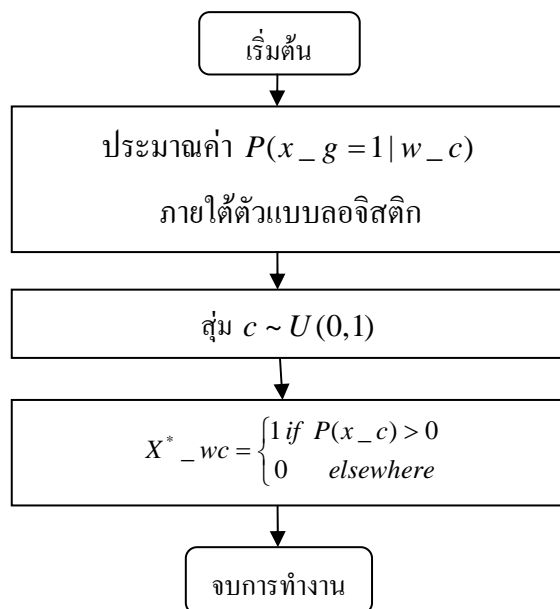
ขั้นตอนของเทคนิคการสอบเทียบลอจิสต์ 1-3 แสดงไว้ในภาพที่ 6



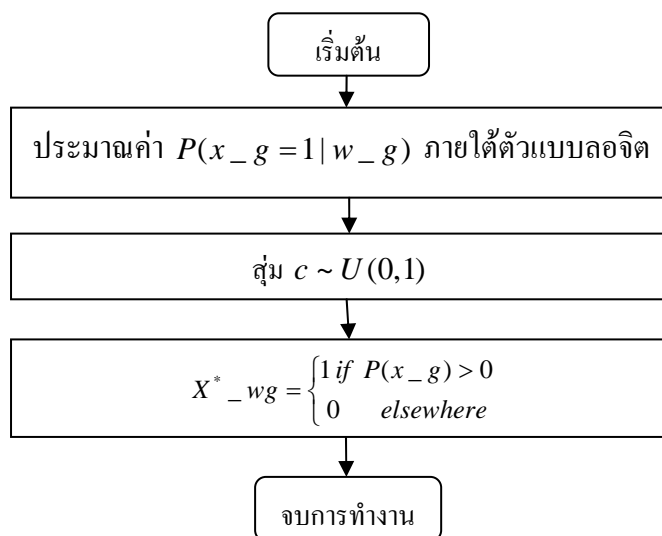
ภาพที่ 3 ขั้นตอนของเทคนิคการสอบเทียบโพรบิต



มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์
ภาพที่ 4 ขั้นตอนของเทคนิคการสอบเทียบคอมพลิเมนต์ารี ล็อก-ล็อก



ภาพที่ 5 ขั้นตอนของเทคนิคการสอบเทียบลอจิสติก



ภาพที่ 6 ขั้นตอนของเทคนิคการสอบเทียบลอจิสติก

2. งานวิจัยที่เกี่ยวข้อง

Reade-Christopher and Kupper (1991) ได้เสนอตัวแบบการถดถอยลอจิสติกและการถดถอยปัวซองในการประเมินผลกระทบของตัวแปรตอบอธิบายที่มีความผิดพลาดจากการจำแนกกลุ่มในการศึกษาติดตามผล วิธีที่นำเสนอมีประโยชน์ต่อผลลัพธ์ในสถานการณ์ที่ต้องใช้การสุ่มตัวอย่างแบบ 2 ชั้นแต่ไม่สามารถทำได้ เนื่องจากความไม่สมบูรณ์ของตัวแปรอธิบาย เช่น การศึกษาทางคลินิกวิทยาต่าง ๆ วิธีดังกล่าวคือการสร้างตัวแบบความเสี่ยงด้วยตัวแบบลอจิสติกและตัวแบบอัตราเสี่ยงและประมาณพารามิเตอร์ที่สนใจด้วยวิธี Weighted least squares หรือเทคนิค Iteratively reweighted maximum likelihood

Schafer (1993) ได้เสนอการวิเคราะห์ภาวะน่าจะเป็น (Likelihood analysis) สำหรับตัวแบบการถดถอยโพบริต เมื่อมีตัวแปรอธิบายบางตัวแบบต่อเนื่องมีความผิดพลาดจากการวัดค่า การแจกแจงของความผิดพลาดและตัวแปรอธิบายที่ไม่ทราบค่าเมื่อกำหนดตัวแปรอธิบายที่ทราบค่าคือการแจกแจงแบบปกติ การประมาณภาวะน่าจะเป็นสูงสุดสามารถคำนวณได้อย่างถูกต้องจาก EM Algorithm เมื่อพิจารณาจากตัวอย่างที่เป็นอิสระกัน n ตัวของ (r_i, u_i, z_i) เมื่อ r_i คือตัวแปรตอบสนองแบบทวิภาค, u_i คือเวกเตอร์ของตัวแปรอธิบายที่ทราบค่าและไม่มีค่าผิดพลาดของค่า

วัดและ z_i คือ n เวกเตอร์ของค่าวัดในตัวแปรอธิบาย x_i ที่ไม่สามารถสังเกตได้ สมมติให้ (i)
 $P(r_i = 1 | u_i, x_i, z_i) = \Phi(\alpha'u_i + \beta x_i)$ เมื่อ $\Phi(\cdot)$ คือฟังก์ชันการแจกแบบปกติมาตรฐาน

Thoresen and Laake (2000) เสนอผลการเปรียบเทียบวิธีการประมาณพารามิเตอร์ 4 วิธี สำหรับตัวแบบความผิดพลาดของค่าวัดในการถดถอยลอจิสติกได้แก่วิธี Regression Calibration, วิธีภาวะน่าจะเป็นสูงสุดโพรบิตแบบประมาณการ (Approximation), วิธีภาวะน่าจะเป็นสูงสุดโดยตรง (Exact) ภายใต้วแบบลอจิสติกและวิธีตัวประมาณแบบง่าย (Naïve estimator) วิธีการต่างๆ ที่เปรียบเทียบจะพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE), ค่าเอนเอียง (Empirical bias) และค่า Coverage เมื่อตัวแปรอธิบายสำหรับลอจิสติกมีความผิดพลาดของค่าวัด ศึกษาโดยการกำหนดให้ตัวแปรอธิบายแบบต่อเนื่อง 2 ตัวที่มีความผิดพลาดของค่าวัด (X) และไม่มีผิดพลาดของค่าวัด (Z) จากการสุ่มตัวอย่างขนาด 1500 และ 150 ภายใต้ว 5 สถานการณ์คือ 1. สถานการณ์ที่เกิดขึ้นจริงในการศึกษาทางคลินิกวิทยาและระบาดวิทยา ($\beta_0 = -2.25, \beta_1 = 0.371, \beta_2 = 0.371$) 2. สถานการณ์ที่พบว่าความผิดพลาดของค่าวัดใหญ่ (σ_v^2) 3. สถานการณ์ที่มีจำนวนของ $Y = 1$ มาก 4. สถานการณ์ที่ได้รับผลจากตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดของค่าวัด (X) มาก 5. สถานการณ์ที่ตัวแปรอธิบายมีความผิดพลาดของค่าวัด (X) ไม่มีการแจกแจงแบบปกติ ($X \sim \chi_1^2$) จากผลลัพธ์ที่ได้ทั้งหมดผู้วิจัยสรุปได้ว่าวิธีที่ควรเลือกใช้เพื่อปรับแก้ความผิดพลาดของค่าวัดคือวิธีการสอบเทียบการถดถอย

Gustafson and Le (2002) ทำการเปรียบเทียบความเอนเอียงภายใต้ตัวแบบการถดถอยอย่างง่ายที่เกิดจากความผิดพลาดของค่าวัดในตัวแปรอธิบายแบบต่อเนื่องและความผิดพลาดจากการจำแนกกลุ่มของตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่ม รวมทั้งความผิดพลาดจากการจำแนกกลุ่มของตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่มที่มาจากตัวแปรอธิบายแบบต่อเนื่องโดยใช้ตัวแปรอิสระครั้งละ 1 ตัว พบว่าการแบ่งกลุ่มลดความเอนเอียงที่เกิดจากการวัดผิด

Buzas, Tosteson and Stefanski (2003) ได้ทบทวนวรรณกรรมเกี่ยวกับความผิดพลาดของการวัดและปัญหาของการสร้างตัวแบบที่เหมาะสม เช่น ตัวแบบที่เกี่ยวข้องกับการเกิดโรคของตัวแปรตอบสนอง Y กับตัวแปรอธิบาย X แบบต่อเนื่องและตัวแปรอธิบาย Z แบบต่อเนื่องที่ไม่มีผิดพลาดของค่าวัด เมื่อทราบค่าของตัวแปร W แบบต่อเนื่อง นอกเหนือจากตัวแปรตอบสนอง Y และตัวแปรอธิบาย Z แบบต่อเนื่อง แม้ว่า การเกิดโรคอาจมีจากความผิดพลาดของค่าวัดแต่

งานวิจัยนี้มุ่งเน้นความผิดพลาดของค่าวัดในตัวแปรอธิบายแบบต่อเนื่อง สิ่งสำคัญสำหรับปัญหาความผิดพลาดของค่าวัดคือปัญหาการประมาณพารามิเตอร์ที่สนใจเมื่อทราบสารสนเทศที่ได้จากตัวอย่างเพียงแค่ Y, Z, W เท่านั้น ดังนั้นสารสนเทศเพิ่มเติมจึงขึ้นอยู่กับชนิดของตัวแบบความผิดพลาดที่สนใจและสิ่งสำคัญในการศึกษาความผิดพลาดของค่าวัดก็คือการวางแผนสำหรับศึกษาที่สามารถนำมาปรับใช้กับข้อมูลที่มีอยู่และให้ผลสรุปที่ถูกต้องแม่นยำ

Weller et al. (2007) ได้พัฒนาวิธีที่สามารถจัดการกับปัญหาความเอนเอียงในค่าประมาณพารามิเตอร์ของการถดถอยที่เกิดจากความผิดพลาดจากการวัดค่าของตัวแปรอธิบายแบบต่อเนื่อง ภายใต้ข้อสมมติของการมีตัวแปรแทน (Surrogates) หลายตัวและใช้วิธีการสอบเทียบการถดถอยเพื่อปรับค่าประมาณของตัวแปรอธิบายแบบต่อเนื่องมีความผิดพลาดจากการวัดค่าข้อเสนอนี้ถูกนำมาใช้ในการศึกษาปัจจัยที่ส่งผลกระทบต่อการเป็นมะเร็งปอดที่สัมพันธ์กับอาชีพและพฤติกรรมของผู้ป่วย

Lyles and Lin (2009) ได้เสนอการวิเคราะห์ Sensitivity ภายใต้ตัวแบบลอจิสติก ศึกษาผลกระทบของความเอนเอียงจากความผิดพลาดจากการจำแนกกลุ่มในการวิเคราะห์ข้อมูล โดยใช้เกณฑ์ของ Sensitivity เพื่อประมาณความน่าจะเป็นในการจำแนกกลุ่มผิด กรณีตัวแปรตอบสนองมีความผิดพลาดของการจำแนกกลุ่มพบว่า การวิเคราะห์ด้วยวิธีนี้จะน่าจะเป็นมีความเหมาะสมที่สุดแต่ในกรณีของตัวแปรอธิบายมีความผิดพลาดจากการจำแนกกลุ่ม ผู้วิจัยนำค่าของ Sensitivity และ Specificity มาใช้สำหรับการสร้างตัวแบบที่สนใจ

บทที่ 3

วิธีดำเนินการวิจัย

1. วิธีการวิจัย

งานวิจัยนี้เสนอเทคนิคการสอบเทียบ 4 วิธีคือเทคนิคการสอบเทียบโพรบิต เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก เทคนิคการสอบเทียบลอจิสติกและเทคนิคการสอบเทียบลอจิต เพื่อปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบายเชิงกลุ่ม 2 กลุ่ม ที่มาจากตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดของค่าวัดสำหรับตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลีเมนทารี ล็อก-ล็อก การศึกษาโดยการจำลองแบบข้อมูล 1,000 ครั้ง ในแต่ละเงื่อนไขของพารามิเตอร์และขนาดตัวอย่าง

2. ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่นำมาใช้ในการวิจัยเป็นข้อมูลจากการจำลองแบบ เมื่อกำหนดตัวแปรอธิบาย X_g แบบเชิงกลุ่ม 2 กลุ่มจากตัวแปรอธิบาย X_c แบบต่อเนื่อง, ตัวแปร W_c แบบต่อเนื่องที่มีความผิดพลาดจากการวัดโดย $W_c = X_c + U$, ตัวแปร W_g แบบเชิงกลุ่ม 2 กลุ่มจากตัวแปร W_c แบบต่อเนื่อง และความผิดพลาดจากการวัด U มีการแจกแจงแบบปกติค่าเฉลี่ย 0 ความแปรปรวน σ_U^2 ภายใต้เงื่อนไขของการจำลองแบบข้อมูล 9 เงื่อนไข σ_U^2 เท่ากับ (0.75, 1 และ 3) และขนาดตัวอย่าง (100, 500 และ 1,000)

3. การประมวลผลของการจำลองแบบ

พัฒนาโปรแกรมโดยการสร้าง Macro program ให้ประมวลผลร่วมกับโปรแกรมสำเร็จรูป SAS[®] version 9.0

4. สถิติที่ใช้ในการวิจัย

ในการศึกษาเทคนิคการปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่มที่มาจากตัวแปรอธิบาย W_c แบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่า โดยการจำลองแบบข้อมูลภายใต้ตัวแบบลอจิสต์ ตัวแบบโพรบิตและตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก พิจารณาความถูกต้องของการประมาณพารามิเตอร์ โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของค่าประมาณพารามิเตอร์ จากการวิเคราะห์ข้อมูลจำลอง M ครั้ง สูตรการคำนวณ MSE มีดังนี้

$$MSE = \frac{\sum_{m=1}^M (\hat{\theta}_m - \theta)^2}{M}$$

เมื่อ θ แทนพารามิเตอร์ความชันของตัวแบบเชิงเส้นน้อยทั่วไปและ $M=1,000$

5. วิธีการจำลองแบบข้อมูล

วิธีการจำลองแบบข้อมูลประกอบด้วย 3 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 สร้างตัวแปรอธิบาย

1. สุ่มตัวแปรอธิบาย X_c แบบต่อเนื่องจากการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 0

ความแปรปรวนเท่ากับ 1

2. สุ่มตัวแปรความผิดพลาดจากการวัด U แบบต่อเนื่องจากการแจกแจงแบบปกติ ค่าเฉลี่ย 0

ความแปรปรวนเท่ากับ σ_U^2

3. สร้างตัวแปรอธิบาย X_g แบบเชิงกลุ่ม 2 กลุ่ม จากตัวแปรอธิบาย X_c แบบต่อเนื่อง โดยที่ $X_g = 1$ ถ้า $X_c > 0$ และ $X_g = 0$ กรณีอื่นๆ

4. สร้างตัวแปร W_c แบบต่อเนื่องและมีความผิดพลาดจากการวัดค่า โดยที่ $W_c = X_c + U$

5. สร้างตัวแปร W_g แบบเชิงกลุ่ม 2 กลุ่ม จากตัวแปร W_c แบบต่อเนื่อง โดยที่ $W_g = 1$ ถ้า $W_c > 0$ และ $W_g = 0$ กรณีอื่นๆ

6. คำนวณค่า $P(x_g = 1 | w_c)$ และ $P(x_g = 1 | w_g)$ จากเทคนิคการสอบเทียบ 4 วิธี โดยใช้ตัวแปรอธิบายคือ W_c และ W_g ตามสมการดังนี้

เทคนิคการสอบเทียบโพรบิต

$$\Phi^{-1}[P(x_{-g} = 1 | w_{-c})] = \beta_0 + \beta_1 w_{-c}$$

หรือ

$$\Phi^{-1}[P(x_{-g} = 1 | w_{-g})] = \beta_0 + \beta_1 w_{-g}$$

เทคนิคการสอบเทียบคอมพลีเมนต์ลอจิสติก ล็อก-ล็อก

$$\log[-\log(1 - P(x_{-g} = 1 | w_{-c}))] = \beta_0 + \beta_1 w_{-c}$$

หรือ

$$\log[-\log(1 - P(x_{-g} = 1 | w_{-g}))] = \beta_0 + \beta_1 w_{-g}$$

เทคนิคการสอบเทียบลอจิสติก

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

เทคนิคการสอบเทียบลอจิต

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

7. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ

Cutpoint

8. สร้างตัวแปร X^*_{-wc} โดยที่ $X^*_{-wc} = 1$ ถ้า $P(x_{-g} = 1 | w_{-c}) > c$ และ $X^*_{-wc} = 0$

กรณีอื่น ๆ และสร้างตัวแปร X^*_{-wg} โดยที่ $X^*_{-wg} = 1$ ถ้า $P(x_{-g} = 1 | w_{-g}) > c$ และ $X^*_{-wg} = 0$

กรณีอื่น ๆ

ขั้นตอนที่ 2 การจำลองตัวแปรสุ่ม Y

1. คำนวณค่า $P(x_{-g})$ เพื่อใช้ในการจำลองตัวแปรสุ่ม Y โดยการแทนค่าพารามิเตอร์

$\beta_0 = -2.25$, $\beta_1 = 0.371$ และค่าของตัวแปร X_{-g} ภายใต้ตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลีเมนต์ลอจิสติก ล็อก-ล็อก ตามสมการดังนี้

ตัวแบบลอจิต

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

ตัวแบบโพรบิต

$$\Phi^{-1}[P(x_g)] = \beta_0 + \beta_1 X_g$$

ตัวแบบคอมพลีเมนต์ารี ล็อก-ล็อก

$$\log[-\log(1 - P(x_g))] = \beta_0 + \beta_1 X_g$$

2. สุ่มค่า c จากการแจกแจงแบบยูนิฟอร์ม ค่าเฉลี่ยเท่ากับ 0.5 เพื่อเป็นเกณฑ์ในการ

Cutpoint

3. สร้างตัวแปรสุ่ม Y พิจารณาจาก $P(x_g)$ ที่ได้จากขั้นตอนที่ 1 โดยที่ $Y = 1$ ถ้า $P(x_g) > c$ และ $Y = 0$ กรณีอื่น ๆ

4. ทำซ้ำขั้นตอนที่ 2-3 จำนวน 1,000 ครั้ง

ขั้นตอนที่ 3 การวิเคราะห์ข้อมูล

1. นำค่าของตัวแปรสุ่ม Y มาทำการวิเคราะห์ภายใต้ตัวแบบลอจิสต์ ตัวแบบโพรบิตและตัวแบบคอมพลีเมนต์ารี ล็อก-ล็อก โดยใช้ตัวแปรอธิบายคือ X_{wc}^* และ X_{wg}^* ตามสมการดังนี้

ตัวแบบลอจิสต์

$$\log \left[\frac{P(Y=1|x_{wc}^*)}{1 - P(Y=1|x_{wc}^*)} \right] = \beta_0 + \beta_1 x_{wc}^*$$

$$\log \left[\frac{P(Y=1|x_{wg}^*)}{1 - P(Y=1|x_{wg}^*)} \right] = \beta_0 + \beta_1 x_{wg}^*$$

ตัวแบบโพรบิต

$$\Phi^{-1}[P(Y=1|x_{wc}^*)] = \beta_0 + \beta_1 x_{wc}^*$$

$$\Phi^{-1}[P(Y=1|x_{wg}^*)] = \beta_0 + \beta_1 x_{wg}^*$$

ตัวแบบคอมพลีเมนต์ารี ล็อก-ล็อก

$$\log[-\log(1 - P(Y=1|x_{wc}^*))] = \beta_0 + \beta_1 x_{wc}^*$$

$$\log[-\log(1 - P(Y=1|x_{wg}^*))] = \beta_0 + \beta_1 x_{wg}^*$$

2. บันทึกค่าประมาณพารามิเตอร์ β_1 ที่ได้ (เฉพาะพารามิเตอร์ที่แสดงอิทธิพลของตัวแปรอธิบายเท่านั้น)

3. ทำซ้ำขั้นตอนที่ 1-2 จำนวน 1,000 ครั้ง

4. คำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณพารามิเตอร์ β_1 ที่ได้ในแต่ละตัวแบบ จากผลการวิเคราะห์ข้อมูลจำลองแบบ 1,000 ครั้ง

ขั้นตอนการจำลองแบบข้อมูลทั้ง 3 ขั้นตอน แสดงไว้ในภาพที่ 7

6. วิธีการวิเคราะห์ข้อมูล

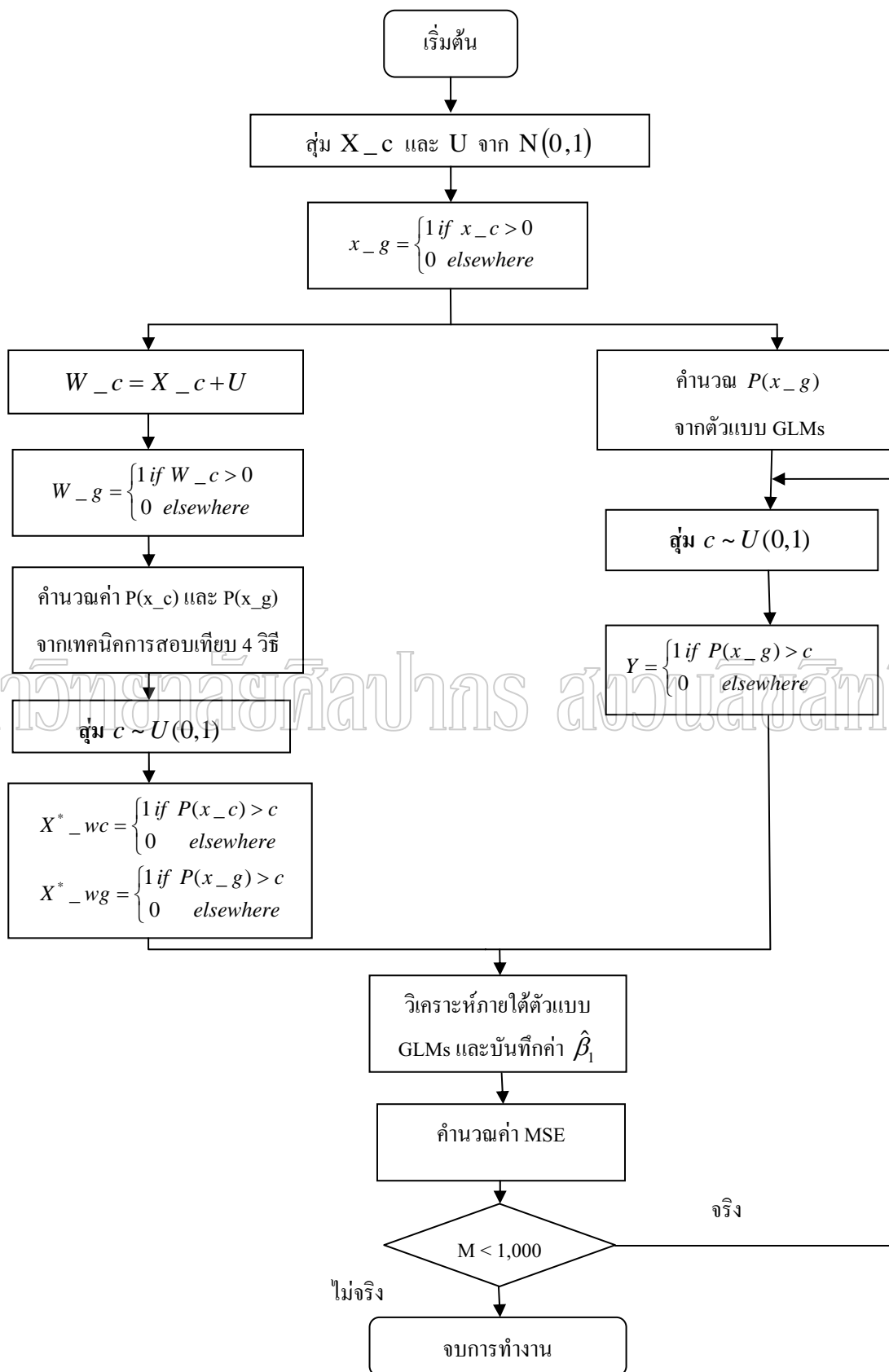
เปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณพารามิเตอร์ที่แสดงอิทธิพลของตัวแปรอธิบายจากเทคนิคการสอบเทียบ 4 วิธี ภายใต้วแบบ GLMs 3 ตัวแบบ โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่น้อยที่สุด

7. การกำหนดความผิดพลาดในการวิจัย

1. ลักษณะของตัวแบบความผิดพลาดของค่าวัดคือ Structural model
2. ความผิดพลาดที่เกิดขึ้นในตัวแปร W คือ Nondifferential error
3. ตัวแบบความผิดพลาดของค่าวัดคือ Classical error models คือ $W_c = X_c + U$

เมื่อ U คือความผิดพลาดที่เกิดขึ้นมีค่าเฉลี่ยเท่ากับ 0 และเป็นอิสระกับ X_c

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์



ภาพที่ 7 ขั้นตอนการจำลองแบบข้อมูล

บทที่ 4

ผลการวิจัย

ในงานวิจัยนี้มุ่งเน้นการนำเสนอเทคนิคการสอบเทียบ 4 วิธี คือ เทคนิคการสอบเทียบโพรบิต (Probit calibration) เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก (Complementary log-log calibration) เทคนิคการสอบเทียบลอจิสติก (Logistic calibration) และเทคนิคการสอบเทียบลอจิต (Logit calibration) เพื่อปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม ที่เกิดจากค่าของตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่า ซึ่งส่งผลกระทบต่อค่าพารามิเตอร์

การจำลองแบบข้อมูลสำหรับตัวแปรตอบสนอง Y และตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่มที่เกิดจากค่าของตัวแปรอธิบาย W_c แบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่าประมาณกลุ่มจริงของตัวแปรอธิบาย X_g ด้วยตัวแปร W_c และตัวแปร W_g จากเทคนิคการสอบเทียบ 4 วิธี เพื่อหาค่าของตัวแปร X^*_{wc} และตัวแปร X^*_{wg} และนำไปใช้ในการวิเคราะห์ภายใต้ตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลีเมนทารี ล็อก-ล็อก สำหรับความแปรปรวนของความผิดพลาด σ_u^2 คือ (0.75, 1 และ 3) และขนาดตัวอย่างคือ (100, 500 และ 1,000) ภายใต้การจำลองข้อมูล 1,000 ครั้งและประมวลผลโดยการสร้าง Macro program ให้ประมวลผลร่วมกับโปรแกรมสำเร็จรูป SAS version 9.0

ผลการวิเคราะห์ข้อมูลสำหรับเทคนิคการสอบเทียบ 4 วิธี เพื่อปรับแก้ความผิดพลาดจากการจำแนกกลุ่มในตัวแปรอธิบาย ด้วยตัวแปร W_c และตัวแปร W_g และผลกระทบจากตัวแบบ GLMs 3 ตัวแบบ คือ ตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลีเมนทารี ล็อก-ล็อก เมื่อตัวแปรอธิบายมีความผิดพลาดจากการจำแนกกลุ่ม โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองที่น้อยที่สุด แสดงตามลำดับดังนี้

ตารางที่ 1 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลอจิต ภายใต้ $\sigma_U^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบลอจิต				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
0.75	100	Probit	0.20472	0.22414
		Comp-log-log	0.22776	0.20601
		Logistic	0.18961	-
		Logit	-	0.22069
	500	Probit	0.02908	0.02762
		Comp-log-log	0.03038	0.02963
		Logistic	0.02769	-
		Logit	-	0.02717
	1,000	Probit	0.01648	0.01515
		Comp-log-log	0.01560	0.01648
		Logistic	0.01548	-
		Logit	-	0.01511

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่น้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 1 พบว่าเมื่อ $\sigma_U^2 = 0.75$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และ 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก สำหรับขนาดตัวอย่าง 100 และเทคนิคการสอบเทียบลอจิต สำหรับขนาดตัวอย่าง 500 และ 1,000

ตารางที่ 2 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลอจิต ภายใต้ $\sigma_U^2 = 1$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบลอจิต				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
1	100	Probit	0.20370	0.23705
		Comp-log-log	0.23694	0.19478
		Logistic	0.20344	
		Logit		0.29436
	500	Probit	0.03033	0.02799
		Comp-log-log	0.03403	0.03151
		Logistic	0.03037	
		Logit	-	0.02999
	1,000	Probit	0.01632	0.01550
		Comp-log-log	0.01559	0.01644
		Logistic	0.01630	-
		Logit	-	0.01559

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 2 พบว่าเมื่อ $\sigma_U^2 = 1$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100 เทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 500 และเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก สำหรับขนาดตัวอย่าง 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก สำหรับขนาดตัวอย่าง 100 และเทคนิคการสอบเทียบลอจิต สำหรับขนาดตัวอย่าง 500 และ 1,000

ตารางที่ 3 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบลอจิต ภายใต้ $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบลอจิต				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
3	100	Probit	0.30170	0.24185
		Comp-log-log	0.24450	0.29392
		Logistic	0.23261	-
		Logit	-	0.26417
	500	Probit	0.03011	0.02905
		Comp-log-log	0.03467	0.03135
		Logistic	0.02936	-
		Logit	-	0.02938
	1,000	Probit	0.01607	0.01579
		Comp-log-log	0.01607	0.01675
		Logistic	0.01591	-
		Logit	-	0.01620

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 3 พบว่าเมื่อ $\sigma_U^2 = 3$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และ 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบลอจิต คือ เทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 100, 500 และ 1,000

จากผลลัพธ์ในตารางที่ 1-3 แสดงให้เห็นว่าภายใต้ตัวแบบลอจิต เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิสติก สำหรับทุก ๆ ขนาด

ตัวอย่างและทุก ๆ ขนาดความแปรปรวนของความผิดพลาด และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิต สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนไม่มาก และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนที่มากขึ้น

ตารางที่ 4 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต้ $\sigma_u^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบโพรบิต				
σ_u^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
0.75	100	Probit	0.07983	0.08762
		Comp-log-log	0.08886	0.08029
		Logistic	0.07389	-
		Logit	-	0.08613
	500	Probit	0.01138	0.01081
		Comp-log-log	0.01189	0.01160
		Logistic	0.01084	-
		Logit	-	0.01063
	1,000	Probit	0.00646	0.00594
		Comp-log-log	0.00611	0.00646
		Logistic	0.00607	-
		Logit	-	0.00592

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 4 พบว่าเมื่อ $\sigma_U^2 = 0.75$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และ 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบลอจิต สำหรับตัวอย่างขนาด 500 และ 1,000

ตารางที่ 5 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต้ $\sigma_U^2 = 1$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบโพรบิต				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
1	100	Probit	0.07958	0.09222
		Comp-log-log	0.09232	0.07604
		Logistic	0.07944	-
		Logit	-	0.11495
	500	Probit	0.01187	0.01096
		Comp-log-log	0.01332	0.01234
		Logistic	0.01189	-
		Logit	-	0.01174
	1,000	Probit	0.00640	0.00607
		Comp-log-log	0.00611	0.00644
		Logistic	0.00639	-
		Logit	-	0.00611

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่น้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 5 พบว่าเมื่อ $\sigma_U^2 = 1$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 500 และเทคนิคคอมพลีเมนทารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 1,000 เทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบคอมพลีเมนทารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 500 และ 1,000

ตารางที่ 6 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบโพรบิต ภายใต้ $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบโพรบิต				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
3	100	Probit	0.11790	0.09440
		Comp-log-log	0.09533	0.11482
		Logistic	0.09056	-
		Logit	-	0.10317
	500	Probit	0.01178	0.01137
		Comp-log-log	0.01356	0.01227
		Logistic	0.01149	-
		Logit	-	0.01149
	1,000	Probit	0.00630	0.00619
		Comp-log-log	0.00630	0.00656
		Logistic	0.00624	-
		Logit	-	0.00635

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 6 พบว่าเมื่อ $\sigma_U^2 = 3$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และ 1,000 เทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบโพรบิต คือ เทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 100, 500 และ 1,000

จากผลลัพธ์ในตารางที่ 4-6 แสดงให้เห็นว่าภายใต้ตัวแบบโพรบิต เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิสติก สำหรับทุก ๆ ขนาดตัวอย่างและทุก ๆ ขนาดความแปรปรวนของความผิดพลาด และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนไม่มาก และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนที่มากขึ้น

ตารางที่ 7 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลีเมนต์รี ล็อก-ล็อก ภายใต้ $\sigma_U^2 = 0.75$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

มหาวิทยาลัยศิลปากร ส่วนวนลิขสิทธิ์
ตัวแบบคอมพลีเมนต์รี ล็อก-ล็อก

σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
0.75	100	Probit	0.10804	0.11816
		Comp-log-log	0.11995	0.11015
		Logistic	0.10140	-
		Logit	-	0.11618
	500	Probit	0.01438	0.01359
		Comp-log-log	0.01500	0.01460
		Logistic	0.01363	-
		Logit	-	0.01338

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

ตารางที่ 7 (ต่อ)

ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
0.75	1,000	Probit	0.00822	0.00753
		Comp-log-log	0.00777	0.0082
		Logistic	0.00772	-
		Logit	-	0.00752

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 7 พบว่าเมื่อ $\sigma_U^2 = 0.75$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และ 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบคอมพลิเมนต์ารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบลอจิต สำหรับตัวอย่างขนาด 500 และ 1,000

ตารางที่ 8 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อกภายใต้ $\sigma_U^2 = 1$ จำนวนตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและแบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
1	100	Probit	0.10769	0.12416
		Comp-log-log	0.12216	0.10386
		Logistic	0.10687	-
		Logit	-	0.15469

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

ตารางที่ 8 (ต่อ)

ตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
1	500	Probit	0.01498	0.01383
		Comp-log-log	0.0168	0.01553
		Logistic	0.01493	-
		Logit	-	0.0148
	1,000	Probit	0.00811	0.00771
		Comp-log-log	0.00776	0.00818
		Logistic	0.00811	-
		Logit	-	0.00778

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 8 พบว่าเมื่อ $\sigma_U^2 = 1$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 100, 500 และเทคนิคการสอบเทียบคอมพลิเมนต์ารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 1,000 และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบคอมพลิเมนต์ารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบคอมพลิเมนต์ารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 500 และ 1,000

ตารางที่ 9 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบคอมพลิเมนทารี ล็อก-ล็อก ภายใต้ $\sigma_U^2 = 3$ จำแนกตามขนาดตัวอย่างและเทคนิคการสอบเทียบ 4 วิธีด้วยตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม

ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก				
σ_U^2	ขนาดตัวอย่าง	เทคนิคการสอบเทียบ	ตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
3	100	Probit	0.15900	0.12944
		Comp-log-log	0.13113	0.15436
		Logistic	0.12437	-
		Logit	-	0.14014
	500	Probit	0.01481	0.01431
		Comp-log-log	0.01699	0.01544
		Logistic	0.01448	-
		Logit	-	0.01440
	1,000	Probit	0.00799	0.00787
		Comp-log-log	0.00801	0.00834
		Logistic	0.00792	-
		Logit	-	0.00807

ตัวเข้ม หมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่น้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 9 พบว่าเมื่อ $\sigma_U^2 = 3$ เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสมสำหรับตัวแบบคอมพลิเมนทารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก สำหรับตัวอย่างขนาด 100 และเทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาด 500 และ 1,000 เทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสมสำหรับตัวแบบคอมพลิเมนทารี ล็อก-ล็อก คือ เทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาด 100, 500 และ 1,000

จากผลลัพธ์ในตารางที่ 7-9 แสดงให้เห็นว่าภายใต้ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก เทคนิคการสอบเทียบด้วยตัวแปร W แบบต่อเนื่องที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิสติก สำหรับทุก ๆ ขนาดตัวอย่างและทุก ๆ ขนาดความแปรปรวนของความผิดพลาด และเทคนิคการสอบเทียบด้วยตัวแปร W แบบ 2 กลุ่มที่เหมาะสม คือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนไม่มาก และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนที่มากขึ้น

จากผลลัพธ์ในตารางที่ 1-9 แสดงให้เห็นถึงเทคนิคการสอบเทียบที่ดีที่สุดภายใต้ตัวแบบลอจิสติก ตัวแบบโพรบิตและตัวแบบคอมพลิเมนทารี ล็อก-ล็อก ดังนั้นสามารถเลือกเทคนิคการสอบเทียบและตัวแบบที่เหมาะสมที่สุด ดังนี้

ตารางที่ 10 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมที่สุดในแต่ละเงื่อนไข ภายใต้ตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม $\sigma_v^2 = 0.75$ ขนาดตัวอย่างและตัวแบบ GLMs

σ_v^2	ขนาดตัวอย่าง	ตัวแบบ	เทคนิคการสอบเทียบที่เหมาะสมด้วยตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
0.75	100	Probit	Logistic(0.07389)	Comp-log-log(0.08029)
		Comp-log-log	Logistic(0.10414)	Comp-log-log(0.11015)
		Logit	Logistic(0.18961)	Comp-log-log(0.20601)
	500	Probit	Logistic(0.01084)	Logit(0.01063)
		Comp-log-log	Logistic(0.01363)	Logit(0.01338)
		Logit	Logistic(0.02769)	Logit(0.02717)
	1,000	Probit	Logistic(0.00607)	Logit(0.00592)
		Comp-log-log	Logistic(0.00772)	Logit(0.00752)
		Logit	Logistic(0.01548)	Logit(0.01511)

- ค่าในวงเล็บหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจากการเปรียบเทียบเทคนิคการสอบเทียบ 4 วิธี ภายใต้ตัวแบบ GLMs

- ตัวเข้มหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 10 เป็นการเปรียบเทียบเทคนิคการสอบเทียบในแต่ละเงื่อนไข เช่น ภายใต้อัตราความคลาดเคลื่อน $\sigma_U^2 = 0.75$, ขนาดตัวอย่างเท่ากับ 100 โดยเลือกมา 1 เทคนิคการสอบเทียบและ 1 ตัวแบบที่เหมาะสมที่สุด เช่น เทคนิคการสอบเทียบลอจิสติกที่เหมาะสมที่สุดในบรรดา 4 เทคนิคและตัวแบบโพรบิตที่เหมาะสมที่สุดในบรรดา 3 ตัวแบบ และภายใต้อัตราความคลาดเคลื่อนขนาด 500 และ 1,000 คือเทคนิคการสอบเทียบลอจิสติก ภายใต้อัตราความคลาดเคลื่อนโพรบิต

ตารางที่ 11 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมที่สุดในแต่ละเงื่อนไข ภายใต้อัตราความคลาดเคลื่อน $\sigma_U^2 = 1$ ขนาดตัวอย่างและตัวแบบ GLMs

σ_U^2	ขนาดตัวอย่าง	ตัวแบบ	เทคนิคการสอบเทียบที่เหมาะสมด้วยตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
1	100	Probit	Logistic(0.07944)	Comp-log-log(0.07604)
		Comp-log-log	Logistic(0.10687)	Comp-log-log(0.10386)
		Logit	Logistic(0.20344)	Comp-log-log(0.19478)
	500	Probit	Probit(0.01187)	Probit(0.01096)
		Comp-log-log	Logistic(0.01493)	Probit(0.01383)
		Logit	Probit(0.03033)	Probit(0.02799)
	1,000	Probit	Comp-log-log(0.00611)	Probit(0.00607)
		Comp-log-log	Comp-log-log(0.00776)	Probit(0.00811)
		Logit	Comp-log-log(0.01559)	Probit(0.01550)

- ค่าในวงเล็บหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจากการเปรียบเทียบเทคนิคการสอบเทียบ 4 วิธี ภายใต้อัตราความคลาดเคลื่อน

- ตัวเข้มหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 11 เป็นการเปรียบเทียบเทคนิคการสอบเทียบในแต่ละเงื่อนไข เช่น ภายใต้อัตราความคลาดเคลื่อน $\sigma_U^2 = 1$, ขนาดตัวอย่างเท่ากับ 100 โดยเลือกมา 1 เทคนิคการสอบเทียบและ 1 ตัวแบบที่เหมาะสมที่สุด เช่น เทคนิคการสอบเทียบคอมพลีเมนต์ ล็อก-ล็อกที่เหมาะสมที่สุดในบรรดา 4 เทคนิคและ

ตัวแบบโพรบิตเหมาะสมที่สุดในบรรดา 3 ตัวแบบ และภายใต้ขนาดตัวอย่างขนาด 500 และ 1,000 คือเทคนิคการสอบเทียบโพรบิต ภายใต้ตัวแบบโพรบิต

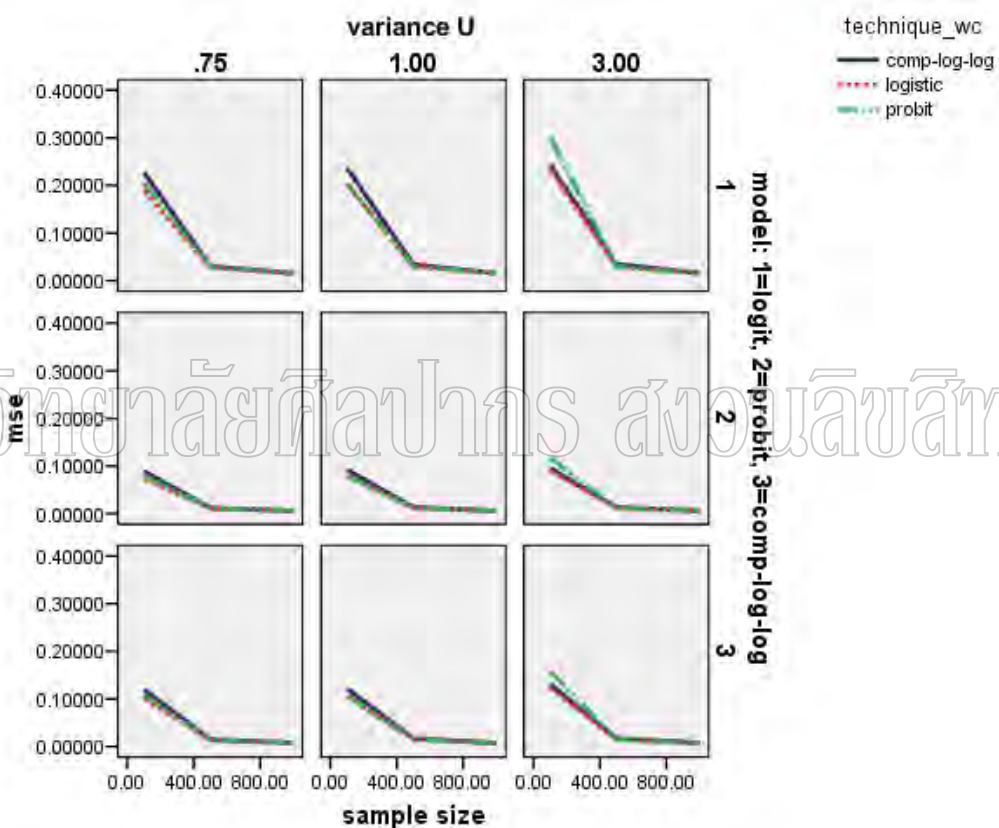
ตารางที่ 12 การเปรียบเทียบเทคนิคการสอบเทียบที่เหมาะสมที่สุดในแต่ละเงื่อนไข ภายใต้ตัวแปร W แบบต่อเนื่องและตัวแปร W แบบเชิงกลุ่ม 2 กลุ่ม $\sigma_U^2 = 3$ ขนาดตัวอย่างและตัวแบบ GLMs

σ_U^2	ขนาดตัวอย่าง	ตัวแบบ	เทคนิคการสอบเทียบที่เหมาะสมด้วยตัวแปร W	
			ต่อเนื่อง (W_c)	2 กลุ่ม (W_g)
3	100	Probit	Logistic(0.09056)	Probit(0.09440)
		Comp-log-log	Comp-log-log(0.13113)	Probit(0.12944)
		Logit	Logistic(0.23261)	Probit(0.24185)
	500	Probit	Logistic(0.01149)	Probit(0.01137)
		Comp-log-log	Logistic(0.01448)	Probit(0.01431)
		Logit	Logistic(0.02936)	Probit(0.02905)
	1,000	Probit	Logistic(0.00624)	Probit(0.00619)
		Comp-log-log	Logistic(0.00792)	Probit(0.00787)
		Logit	Logistic(0.01591)	Probit(0.01579)

- ค่าในวงเล็บหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจากการเปรียบเทียบเทคนิคการสอบเทียบ 4 วิธี ภายใต้ตัวแบบ GLMs
- ตัวเข้มหมายถึงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดสำหรับแต่ละเงื่อนไข

จากตารางที่ 12 เป็นการเปรียบเทียบเทคนิคการสอบเทียบในแต่ละเงื่อนไข เช่น ภายใต้ $\sigma_U^2 = 3$, ขนาดตัวอย่างเท่ากับ 100 โดยเลือกมา 1 เทคนิคการสอบเทียบและ 1 ตัวแบบที่เหมาะสมที่สุด เช่น เทคนิคการสอบเทียบลอจิสติกเหมาะสมที่สุดในบรรดา 4 เทคนิคและตัวแบบโพรบิตเหมาะสมที่สุดในบรรดา 3 ตัวแบบ และภายใต้ขนาดตัวอย่างขนาด 500 และ 1,000 คือเทคนิคการสอบเทียบโพรบิต ภายใต้ตัวแบบโพรบิต

จากผลลัพธ์ในตารางที่ 10-12 แสดงให้เห็นว่าเทคนิคการสอบเทียบที่เหมาะสมที่สุดคือ เทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาดเล็ก ในทุก ๆ ขนาดความแปรปรวนของความผิดพลาด และเทคนิคการสอบเทียบลอจิสติก สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนไม่มาก และเทคนิคการสอบเทียบโพรบิต สำหรับตัวอย่างขนาดใหญ่และขนาดความแปรปรวนที่มากขึ้น ส่วนตัวแบบที่เหมาะสมที่สุดคือ ตัวแบบโพรบิต สำหรับทุก ๆ ขนาดตัวอย่างและทุก ๆ ขนาดความแปรปรวน

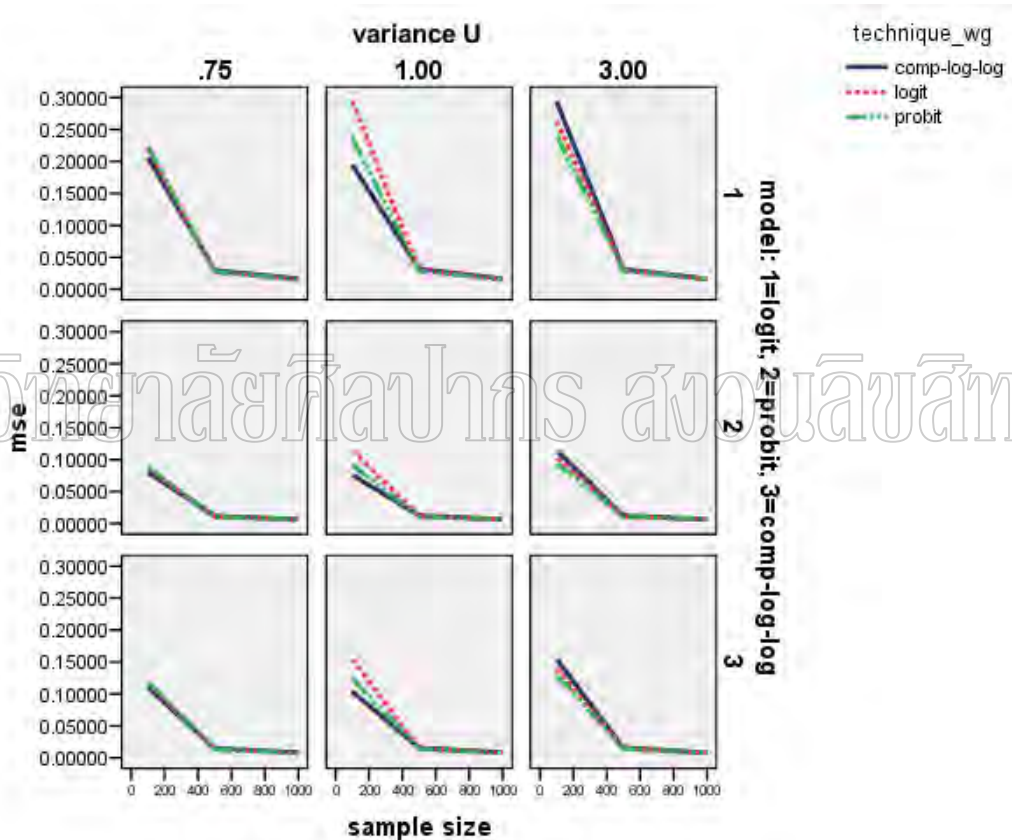


ภาพที่ 8 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จากเทคนิคการสอบเทียบ 3 วิธีที่ปรับแก้ด้วยตัวแปรอธิบาย W_c แบบต่อเนื่อง ภายใต้ความแปรปรวนของความผิดพลาด ขนาดตัวอย่างและตัวแบบ GLMS

จากภาพที่ 8 แสดงให้เห็นว่าเทคนิคการสอบเทียบโพรบิตให้ผลดีพอ ๆ กับเทคนิคการสอบเทียบลอจิสติก แต่เมื่อความแปรปรวนของความผิดพลาดมีขนาดใหญ่เทคนิคการสอบเทียบ

โพรบิตให้ผลแย่กว่าเทคนิคการสอบเทียบลอจิสติกเล็กน้อยและเมื่อขนาดตัวอย่างใหญ่ขึ้นเทคนิคการสอบเทียบทั้ง 3 เทคนิคให้ผลดีพอ ๆ กัน

ตัวแบบที่มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด คือ ตัวแบบโพรบิต รองลงมาคือ ตัวแบบคอมพลีเมนต์ทารี ล็อก-ล็อก และตัวแบบลอจิสติก ดังนั้นตัวแบบที่เหมาะสมที่สุดคือ ตัวแบบโพรบิต



ภาพที่ 9 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จากเทคนิคการสอบเทียบ 3 วิธีที่ปรับแก้ด้วยตัวแปรอธิบาย W_g แบบ 2 กลุ่ม ภายใต้ความแปรปรวนของความผิดพลาด ขนาดตัวอย่างและตัวแบบ GLMS

จากภาพที่ 9 แสดงให้เห็นว่าเทคนิคการสอบเทียบโพรบิตให้ผลดีกว่าเทคนิคการสอบเทียบลอจิสติกซึ่งเห็นอย่างชัดเจนเมื่อความแปรปรวนของความผิดพลาดมีขนาดใหญ่ขึ้น แต่เมื่อขนาด

ตัวอย่างใหญ่ขึ้นเทคนิคการสอบเทียบทั้ง 3 เทคนิคให้ผลดีพอ ๆ กัน ส่วนเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก ให้ผลดีเฉพาะความแปรปรวนของความผิดพลาดมีขนาดไม่มาก

ตัวแบบที่มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด คือ ตัวแบบ โพรบิต รองลงมา คือ ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก และตัวแบบลอจิต ดังนั้นตัวแบบที่เหมาะสมที่สุดคือ ตัวแบบ โพรบิต

ดังนั้นจากภาพที่ 8 และ 9 พบว่าการสอบเทียบด้วยตัวแปรอธิบาย W_g แบบ 2 กลุ่มให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าการสอบเทียบด้วยตัวแปรอธิบาย W_c แบบต่อเนื่อง และเทคนิคการสอบเทียบที่เหมาะสมที่สุดในการปรับแก้ความผิดพลาดคือ เทคนิคการสอบเทียบโพรบิต รองลงมาคือเทคนิคการสอบเทียบลอจิสติกและเทคนิคการสอบเทียบลอจิต และเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

1. สรุปและอภิปรายผลการวิจัย

เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิดในตัวแบบเชิงเส้นน้อยทั่วไป เสนอวิธีปรับแก้ความผิดพลาดที่ไม่ซับซ้อน 4 วิธี คือ การสอบเทียบโพรบิต (Probit calibration) การสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก (Complementary log-log calibration) การสอบเทียบลอจิสติก (Logistic calibration) และการสอบเทียบลอจิด (Logit calibration) การจำลองแบบข้อมูลสำหรับตัวแปรตอบสนองและกำหนดค่าตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่เกิดจากการแบ่งค่าของตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่าสำหรับตัวอย่างขนาด 100, 500 และ 1,000 เมื่อกำหนดให้ความแปรปรวนของความผิดพลาดคือ 0.75, 1 และ 3 ภายใต้ตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิด ตัวแบบโพรบิตและตัวแบบคอมพลิเมนทารี ล็อก-ล็อก ซึ่งเป็นตัวแบบที่มีเทอมหลักคล้ายกัน จากผลลัพธ์ที่พบภายใต้งานวิจัยสามารถสรุปได้ดังนี้

เทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_c แบบต่อเนื่องที่เหมาะสมที่สุด คือ เทคนิคการสอบเทียบโพรบิต ซึ่งให้ผลดีพอ ๆ กับเทคนิคการสอบเทียบลอจิสติก แต่เมื่อความแปรปรวนของความผิดพลาดมีขนาดใหญ่เทคนิคการสอบเทียบโพรบิตให้ผลแย่กว่าเทคนิคการสอบเทียบลอจิสติกเล็กน้อยและเมื่อขนาดตัวอย่างใหญ่ขึ้นเทคนิคการสอบเทียบทั้ง 3 เทคนิคให้ผลดีพอ ๆ กัน

เทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่มที่เหมาะสมที่สุด คือ เทคนิคการสอบเทียบโพรบิต ซึ่งให้ผลดีกว่าเทคนิคการสอบเทียบลอจิดเห็นอย่างชัดเจนเมื่อความแปรปรวนของความผิดพลาดมีขนาดใหญ่ขึ้น ส่วนเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อกให้ผลดีเฉพาะความแปรปรวนของความผิดพลาดมีขนาดไม่มาก และเมื่อขนาดตัวอย่างใหญ่ขึ้นเทคนิคการสอบเทียบทั้ง 3 เทคนิคให้ผลดีพอ ๆ กัน

เมื่อเปรียบเทียบระหว่างเทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_c แบบต่อเนื่องและเทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่ม พบว่าเทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่ม ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าการสอบเทียบด้วยตัวแปรอธิบาย W_c แบบต่อเนื่อง ดังนั้นเทคนิคการสอบเทียบที่เหมาะสมที่สุดในการปรับแก้ความผิดพลาด คือ เทคนิคการสอบเทียบโพรบิต รองลงมาคือเทคนิคการสอบเทียบลอจิสติก เทคนิคการสอบเทียบลอจิตและเทคนิคการสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก

ตัวแบบที่มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด คือ ตัวแบบโพรบิต รองลงมาคือ ตัวแบบคอมพลิเมนทารี ล็อก-ล็อก และตัวแบบลอจิต สำหรับเทคนิคการสอบเทียบด้วยตัวแปรอธิบาย W_c แบบต่อเนื่องและตัวแปรอธิบาย W_g แบบเชิงกลุ่ม 2 กลุ่ม ดังนั้นตัวแบบที่เหมาะสมที่สุดคือ ตัวแบบโพรบิต

2. ข้อเสนอแนะของงานวิจัย

จากผลการวิจัยพบว่าเทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดด้วยตัวแปรอธิบาย W_g ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าการปรับแก้ด้วยตัวแปรอธิบาย W_c และเทคนิคการสอบเทียบที่ควรเลือกใช้คือเทคนิคการสอบเทียบโพรบิต เพื่อพยากรณ์หรือประมาณกลุ่มจริงก่อนการวิเคราะห์และจะทำให้ค่าประมาณพารามิเตอร์ในตัวแบบดีขึ้นเมื่อใช้พยากรณ์ด้วยตัวแปรอธิบาย W_g ส่วนตัวแบบที่ควรเลือกใช้คือตัวแบบโพรบิต

3. ข้อเสนอแนะในงานวิจัยครั้งต่อไป

เนื่องจากงานวิจัยนี้นำเสนอวิธีการปรับแก้ความผิดพลาดจากการวัดค่าในตัวแปรอธิบายตัวแบบเชิงเส้นน้อยเกินไปเพียง 1 ตัวเท่านั้น โดยศึกษาในกรณีที่ไม่ซับซ้อน คือ ตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่ม ทำให้เทคนิคการสอบเทียบสามารถประมาณกลุ่มจริงของตัวแปรอธิบายได้เพียงแค่ 2 กลุ่ม ในงานวิจัยครั้งต่อไปอาจเพิ่มจำนวนตัวแปรอธิบายที่มีความผิดพลาดจากการวัดค่าให้มากกว่า 1 ตัว หรืออาจใช้ตัวแปรอธิบายที่มากกว่า 2 กลุ่มทั้งแบบ Ordinal scale และ Nominal scale และอาจเพิ่มการศึกษาเกี่ยวกับตัวแบบเชิงเส้นน้อยทั่วไปที่พิจารณาให้มากขึ้น รวมทั้งอาจกำหนดให้ความสัมพันธ์ระหว่างตัวแปร W_c และตัวแปร X_c มีรูปแบบที่ซับซ้อนมากขึ้น

บรรณานุกรม

ภาษาไทย

วีรพันธ์ พงศาภักดี. การวิเคราะห์ข้อมูลเชิงกลุ่ม. นครปฐม : โรงพิมพ์มหาวิทยาลัยศิลปากร, 2544.

ภาษาต่างประเทศ

Agresti, A. Categorical data analysis. 2nd ed. New York : John Wiley & Sons, 2002.

Armstrong, B. "Measurement error in generalized linear models." Communications in statistics part B – Simulation and Commutation, no. 14(1985) : 529-544.

Buzas, J. S., T. D Tosteson, and L. A. Stefanski. "Measurement error." Institute of Statistics Mimeo Series, no. 2544(2003) : 1-31.

Carroll, R.J., D. Ruppert, and L. A. Stefanski. Measurement error in nonlinear models. New York : Chapman and Hall, 1995.

Weller, A. et al. "Regression calibration for logistic regression with multiple surrogates for one exposure." Journal of Statistical Planning and Inference, no. 137(2007) : 449-461.

Fuller, W.A. Measurement Error Models. New York : John Wiley & Sons, 1987.

Gustafson, P., and D. Le. Nhu. "Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors." Biometrics 58, 4(December 2002) : 878-887.

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

Nelder, J., and R. W. M. Wedderburn. "Generalized linear models." J. Roy. Statist. Soc., no. 135(1972) : 370-384.

Prentice, R. L. "Covariate measurement errors and parameter estimation in a failure time regression model." Biometrika, no. 69(1982) : 331-342.

- Robert, H. Lyles., and Ji Lin. "Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting." Technical Report 9, 2(July 2009) : 1-31.
- Schafer, D. W. "Likelihood analysis for probit regression with measurement errors." Biometrika 80, 4(December 1993) : 899-904.
- Reade-Christopher, Susan J., and Lawrence L. Kupper. "Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data." Biometric 47, 2(June 1991) : 535-548.
- Thoresen, M., and P. Laake. "A simulation study of measurement error correction methods in logistic regression." Biometrics 56, 3(September 2000) : 868-872.

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ภาคผนวก
มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

โปรแกรมคอมพิวเตอร์ที่ใช้ในงานวิจัย

เทคนิคการสอบเทียบสำหรับการปรับแก้ความผิดพลาดของการวัดและการจำแนกผิดในตัวแบบเชิงเส้นน้อยทั่วไปเสนอวิธีปรับแก้ความผิดพลาดที่ไม่ซับซ้อน 4 วิธีคือการสอบเทียบโพรบิต (Probit calibration) การสอบเทียบคอมพลิเมนทารี ล็อก-ล็อก (Complementary log-log calibration) การสอบเทียบลอจิสติก (Logistic calibration) และการสอบเทียบลอจิต (Logit calibration) จากการจำลองแบบข้อมูลสำหรับตัวแปรตอบสนองและตัวแปรอธิบายแบบเชิงกลุ่ม 2 กลุ่มที่เกิดจากการแบ่งค่าของตัวแปรอธิบายแบบต่อเนื่องที่มีความผิดพลาดจากการวัดค่าสำหรับตัวอย่างขนาด 100, 500 และ 1,000 เมื่อกำหนดให้ความแปรปรวนของความผิดพลาดคือ 0.75, 1 และ 3 ภายใต้ตัวแบบ GLMs 3 ตัวแบบคือตัวแบบลอจิต ตัวแบบโพรบิตและตัวแบบคอมพลิเมนทารี ล็อก-ล็อก โดยแสดงโปรแกรมสำหรับการจำลองข้อมูลและการประมวลผลข้อมูล ตามลำดับ ดังนี้

1. โปรแกรมสำหรับการจำลองแบบข้อมูล

การจำลองแบบข้อมูลสำหรับตัวแปรตอบสนองแบบเชิงกลุ่ม 2 กลุ่ม ตัวแปรอธิบาย X ที่มีการแจกแจงปกติค่าเฉลี่ย 0 ความแปรปรวน 1 ความผิดพลาดที่เกิดขึ้นจากการวัด U มีการแจกแจงปกติค่าเฉลี่ย 0 ความแปรปรวน $\sigma_U^2 = 0.75, 1, 3$ สำหรับตัวอย่างเท่ากับ 100, 500 และ 1,000 ภายใต้การทำซ้ำ 1,000 ครั้ง ซึ่งความหมายของสัญลักษณ์ต่าง ๆ ในตัวโปรแกรม และการเขียนโปรแกรม มีรายละเอียดดังนี้

nrep	หมายถึง	จำนวนครั้งของการจำลองข้อมูล
N	หมายถึง	ขนาดตัวอย่าง
mu	หมายถึง	ค่าเฉลี่ยของความผิดพลาดที่เกิดขึ้นจากการวัด U
sigma	หมายถึง	ความแปรปรวนของความผิดพลาดที่เกิดขึ้นจากการวัด U

```

data sim;
  mu=0;
  sigma=0.75;
  N=100;
  nrep=1000;
  seed=12345;
  do obs = 1 to N;
    X_c = rannor(seed);
    if X_c > 0 then X_g = 1;
    else X_g = 0;
    U = mu+sigma*rannor(seed);
    W_c = X_c + U;
    if W_c > 0 then W_g = 1;
    else W_g = 0;
  do rep = 1 to nrep;
  output;
  end;
  end;
run;
proc sort data=sim;by rep;
run;

```

2. การสร้างโปรแกรมจำลองตัวแปรตอบสนอง Y แบบเชิงกลุ่ม 2 กลุ่ม เมื่อกำหนดค่าพารามิเตอร์

$\beta_0 = -2.25, \beta_1 = 0.371$ (Thoresen and Laake (2000)) ภายใต้ตัวแบบ GLMs และ Cut Point ด้วยการแจกแจงแบบยูนิฟอร์ม (0,1)

```

data p_logit;set sim;
beta0=-2.25;
beta1=0.371;
p_logit = exp(beta0+(beta1*X_g))/1+exp(beta0+(beta1*X_g));
run;
proc print data=p_logit;
run;

title 'cutpoint1';
data cutpoint1;
  nrep=1000;
  N=100;
  seed=12345;
  do rep = 1 to nrep;
  do obs = 1 to N;
    cutpoint1 =1*ranuni(seed)+ 0;
  output;
  end;
  end;
run;
proc print data=cutpoint1;
run;

```



```

data Y_sim_logit;set sim;set cutpoint1;set p_logit;
if p_logit > cutpoint1 then Y = 1;
else Y = 0;
run;
proc print data=Y_sim_logit;
run;

data p_probit;set sim;
beta0=-2.25;
beta1=0.371;
mul=beta0/beta1;
sigma1=1/beta1;
p_probit = CDF('NORMAL',X_g,mul,sigma1);
run;
proc print data=p_probit;
run;

title 'cutpoint2';
data cutpoint2;
  nrep=1000;
  N=100;
  seed=12345;
  do rep = 1 to nrep;
  do obs = 1 to N;
    cutpoint2 =1*ranuni(seed)+ 0;
  output;
  end;
run;
proc print data=cutpoint2;
run;

data Y_sim_probit;set sim;set cutpoint2;set p_probit;
if p_probit > cutpoint2 then Y = 1;
else Y = 0;
run;
proc print data=Y_sim_probit;
run;

data p_clog;set sim;
beta0=-2.25;
beta1=0.371;
p_clog = 1-exp(-exp(beta0+(beta1*X_g)));
run;
proc print data=p_clog;
run;

title 'cutpoint3';
data cutpoint3;
  nrep=1000;N=100;
  seed=12345;
  do rep = 1 to nrep;
  do obs = 1 to N;
    cutpoint3 =1*ranuni(seed)+ 0;
  output;
  end;
end;

```

มหาวิทยาลัยศิลปากร ส่วนวนลิขสิทธิ์

```

run;
proc print data=cutpoint3;run;

data Y_sim_clog;set sim;set cutpoint3;set p_clog;
if p_clog > cutpoint3 then Y = 1;
else Y = 0;
run;
proc print data=Y_sim_clog;
run;

```

3. การสร้างโปรแกรมสำหรับการสร้างเทคนิคการสอบเทียบโพรบิต เพื่อพยากรณ์หรือประมาณกลุ่ม จริงของตัวแปรอธิบาย X^*_wc และ X^*_wg และ Cut Point ด้วยการแจกแจงแบบยูนิฟอร์ม (0,1)

```

title 'Probit Calibration';
proc logistic data=Y_sim_probit;
  model X_g(event='1')=W_c/link=normit;
  ods output parameterestimates=b__normit_calib;
  output out=pred_normit_calib p=p_normit_calib;
run;

```

```

data cutpoint4;
  N=100;seed=12345;
  do obs = 1 to N;
    cutpoint4 = 1*ranuni(seed)+ 0;
  output;
  end;
run;

```

```

data X_wc;set pred_normit_calib;set cutpoint4;
  if p_normit_calib > cutpoint4 then X_wc = 1;
  else X_wc = 0;
  output;
run;

```

```

data X_wc_probit;set X_wc;
  nrep=1000;
  do rep = 1 to nrep;
    if X_wc = 1 then X_wc_probit = 1;
    else X_wc_probit = 0;
  output;
  end;
run;

```

```

proc sort data= X_wc_probit;by rep;run;

```

```

proc logistic data=Y_sim_probit;
  model X_g(event='1')=W_g/link=normit;
  ods output parameterestimates=b__normit_calib;
  output out=pred_normit_calib p=p_normit_calib;
run;

```

```

data cutpoint5;
  N=100;seed=12345;
  do obs = 1 to N;
    cutpoint5 = 1*ranuni(seed)+ 0;

```

```

        output;
        end;
run;
data X_wg;set pred_normit_calib;set cutpoint5;
    if p_normit_calib > cutpoint5 then X_wg = 1;
    else X_wg = 0;
    output;
run;

data X_wg_probit;set X_wg;
    nrep=1000;
    do rep = 1 to nrep;
        if X_wg = 1 then X_wg_probit = 1;
        else X_wg_probit = 0;
    output;
    end;
run;
proc sort data=X_wg_probit;by rep;
run;

```

4. การสร้างโปรแกรมสำหรับการสร้างเทคนิคการสอบเทียบคอมพิวเตอร์ ล็อก-ล็อก เพื่อพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ X^*_wg และ Cut Point ด้วยการแจกแจงแบบยูนิฟอร์ม (0,1)

```

title 'Cloglog Calibration';
proc logistic data=Y_sim_clog;
    model X_g(event='1')=W_c/link=cloglog;
    ods output parameterestimates=b__clog_calib;
    output out=pred_clog_calib p=p_clog_calib;
run;

data cutpoint6;
    N=100;
    seed=12345;
    do obs = 1 to N;
        cutpoint6 = 1*ranuni(seed)+ 0;
    output;
    end;
run;

data X_wc;set pred_clog_calib;set cutpoint6;
    if p_clog_calib > cutpoint6 then X_wc = 1;
    else X_wc = 0;
    output;
run;

data X_wc_clog;set X_wc;
    nrep=1000;
    do rep = 1 to nrep;
        if X_wc = 1 then X_wc_clog = 1;
        else X_wc_clog = 0;
    output;
    end;
run;

```

```

proc sort data= X_wc_clog;by rep;
run;

proc logistic data=Y_sim_clog;
  model X_g(event='1')=W_g/link=cloglog;
  ods output parameterestimates=b__clog_calib;
  output out=pred_clog_calib p=p_clog_calib;
run;

data cutpoint7;
  N=100;
  seed=12345;
  do obs = 1 to N;
    cutpoint7 = 1*ranuni(seed)+ 0;
  output;
  end;
run;

data X_wg;set pred_clog_calib;set cutpoint7;
  if p_clog_calib > cutpoint7 then X_wg = 1;
  else X_wg = 0;
  output;
run;

data X_wg_clog;set X_wg;
  nrep=1000;
  do rep = 1 to nrep;
    if X_wg = 1 then X_wg_clog = 1;
    else X_wg_clog = 0;
  output;
  end;
run;
proc sort data= X_wg_clog;by rep;
run;

```

5. การสร้างโปรแกรมสำหรับการสร้างเทคนิคการสอบเทียบลอจิสติก เพื่อพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ X^*_wg และ Cut Point ด้วยการแจกแจงแบบยูนิฟอร์ม (0,1)

```

title 'Logistic Calibration';
proc logistic data=Y_sim_logit;
  model X_g(event='1')=W_c/link=logit;
  ods output parameterestimates=b__logistic_calib;
  output out=pred_logistic_calib p=p_logistic_calib;
run;

data cutpoint8;
  N=100;
  seed=12345;
  do obs = 1 to N;
    cutpoint8 = 1*ranuni(seed)+ 0;
  output;

```

```

end;
run;

data X_wc;set pred_logistic_calib;set cutpoint8;
  if p_logistic_calib > cutpoint8 then X_wc = 1;
  else X_wc = 0;
output;
run;

data X_wc_logistic;set X_wc;
  nrep=1000;
  do rep = 1 to nrep;
    if X_wc = 1 then X_wc_logistic = 1;
    else X_wc_logistic = 0;
  output;
  end;
run;
proc sort data=X_wc_logistic;by rep;
run;

```

6. การสร้างโปรแกรมสำหรับการสร้างเทคนิคการสอบเทียบลอจิสต์ เพื่อพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ X^*_wg และ Cut Point ด้วยการแจกแจงแบบยูนิฟอร์ม (0,1)

```

title 'Logit Calibration';
proc logistic data=Y_sim_logit;
  model X_g(event='1')=W_g/link=logit;
  ods output parameterestimates=b__logit_calib;
  output out=pred_logit_calib p=p_logit_calib;
run;

data cutpoint9;
  N=100;
  seed=12345;
  do obs = 1 to N;
    cutpoint9 = 1*ranuni(seed)+ 0;
  output;
  end;
run;

data X_wg;set pred_logit_calib;set cutpoint9;
  if p_logit_calib > cutpoint9 then X_wg = 1;
  else X_wg = 0;
output;
run;

data X_wg_logit;set X_wg;
  nrep=1000;
  do rep = 1 to nrep;
    if X_wg = 1 then X_wg_logit = 1;
    else X_wg_logit = 0;
  output;
  end;
run;

```

```
proc sort data= X_wg_logit;by rep;
run;
```

7. การสร้างโปรแกรมสำหรับการนำค่าพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และตัวแปรอธิบาย X^*_wg จากเทคนิคการสอบเทียบโพรบิต เพื่อประมาณค่าพารามิเตอร์และค่าความคลาดเคลื่อนเฉลี่ยกำลังสองภายใต้ตัวแบบ GLMs

```
title 'Logit Model (Y_sim_logit with X*_wc)';
proc logistic data=Y_sim_logit;by rep;
  model Y (event='1')= X_wc_probit /link=logit;
  ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
  var_logit_calib=Variable;
  est_logit_calib=Estimate;
  keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
  if var_logit_calib = ' X_wc_probit ' then do;
    mse_logit_calib=(est_logit_calib - b1)**2;
    output;
  end;
  keep var_logit_calib mse_logit_calib;
run;
proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Logit Model (Y_sim_logit with X*_wg)';
proc logistic data=Y_sim_logit;by rep;
  model Y (event='1')= X_wg_probit /link=logit;
  ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
  var_logit_calib=Variable;
  est_logit_calib=Estimate;
  keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
  if var_logit_calib = ' X_wg_probit ' then do;
```

```

        mse_logit_calib=(est_logit_calib - b1)**2;
        output;
    end;
    keep var_logit_calib mse_logit_calib;
run;

proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Probit Model (Y_sim_probit with X*_wc)';
proc logistic data=Y_sim_probit;by rep;
    model Y (event='1')= X_wc_probit /link=normit;
    ods output parameterestimates=b_probit_calib;
run;

data b_probit_calib;set b_probit_calib;by rep;
    var_probit_calib=Variable;
    est_probit_calib=Estimate;
    keep rep var_probit_calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_probit_calib;set b_probit_calib;
b1= 0.371;
    if var_probit_calib = ' X_wc_probit ' then do;
        mse_probit_calib=(est_probit_calib - b1)**2;
        output;
    end;
    keep var_probit_calib mse_probit_calib;
run;
proc univariate data=b_probit_calib;by var_probit_calib;
var mse_probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;

title 'Probit Model (Y_sim_ probit with X*_wg)';
proc logistic data=Y_sim_ probit;by rep;
    model Y (event='1')= X_wg_probit /link=normit;
    ods output parameterestimates=b_ probit _calib;
run;

data b_probit_calib;set b_probit_calib;by rep;
    var_probit_calib=Variable;
    est_probit_calib=Estimate;
    keep rep var_probit _calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_probit_calib;set b_probit_calib;
b1 = 0.371;
    if var_probit_calib = ' X_wg_probit ' then do;
        mse_ probit_calib=(est_probit_calib - b1)**2;
        output;
    end;

```

```

        end;
        keep var_probit_calib mse_ probit_calib;
run;
proc univariate data=b_probit_calib;by var_probit_calib;
var mse_ probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;
title 'Cloglog Model (Y_sim_clog with X*_wc)';
proc logistic data=Y_sim_clog;by rep;
model Y (event='1')= X_wc_probit /link=cloglog;
ods output parameterestimates=b_clog_calib;
run;

data b_clog_calib;set b_clog_calib;by rep;
var_clog_calib=Variable;
est_clog_calib=Estimate;
keep rep var_clog_calib est_clog_calib;
run;
proc sort data=b_clog_calib;by var_clog_calib;
run;

data b_clog_calib;set b_clog_calib;
b1 = 0.371;
if var_clog_calib = ' X_wc_probit ' then do;
mse_clog_calib=(est_clog_calib - b1)**2;
output;
end;
keep var_clog_calib mse_clog_calib;
run;
proc univariate data=b_clog_calib;by var_clog_calib;
var mse_clog_calib;
output out=mse_clog_calib mean=mse_clog_calib;
run;

title 'Cloglog Model (Y_sim_clog with X*_wg)';
proc logistic data=Y_sim_clog;by rep;
model Y (event='1')= X_wg_probit /link=cloglog;
ods output parameterestimates=b_clog_calib;
run;

data b_clog_calib;set b_clog_calib;by rep;
var_clog_calib=Variable;
est_clog_calib=Estimate;
keep rep var_clog_calib est_clog_calib;
run;
proc sort data=b_clog_calib;by var_clog_calib;
run;

data b_clog_calib;set b_clog_calib;
b1 = 0.371;
if var_clog_calib = ' X_wg_probit ' then do;
mse_clog_calib=(est_clog_calib - b1)**2;
output;
end;
keep var_clog_calib mse_clog_calib;
run;
proc univariate data=b_clog_calib;by var_clog_calib;

```



```
var mse_ clog_calib;
output out=mse_clog_calib mean=mse_clog_calib;
run;
```

8. การสร้างโปรแกรมสำหรับการนำค่าพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ ตัวแปรอธิบาย X^*_wg จากเทคนิคการสอบเทียบคอมพลีเมนต์ทารี ล็อก-ล็อก เพื่อประมาณค่าพารามิเตอร์และค่าความคลาดเคลื่อนเฉลี่ยกำลังสองภายใต้ตัวแบบ GLMs

```
title 'Logit Model (Y_sim_logit with X*_wc)';
proc logistic data=Y_sim_logit;by rep;
  model Y (event='1')= X_wc_clog /link=logit;
  ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
  var_logit_calib=Variable;
  est_logit_calib=Estimate;
  keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
  if var_logit_calib = ' X_wc_clog ' then do;
    mse_logit_calib=(est_logit_calib - b1)**2;
    output;
  end;
  keep var_logit_calib mse_logit_calib;
run;
proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Logit Model (Y_sim_logit with X*_wg)';
proc logistic data=Y_sim_logit;by rep;
  model Y (event='1')= X_wg_clog /link=logit;
  ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
  var_logit_calib=Variable;
  est_logit_calib=Estimate;
  keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
  if var_logit_calib = ' X_wg_clog ' then do;
```

```

        mse_logit_calib=(est_logit_calib - b1)**2;
        output;
    end;
    keep var_logit_calib mse_logit_calib;
run;

proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Probit Model (Y_sim_probit with X*_wc)';
proc logistic data=Y_sim_probit;by rep;
    model Y (event='1')= X_wc_clog /link=normit;
    ods output parameterestimates=b_probit_calib;
run;

data b_probit_calib;set b_probit_calib;by rep;
    var_probit_calib=Variable;
    est_probit_calib=Estimate;
    keep rep var_probit_calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_probit_calib;set b_probit_calib;
b1=0.371;
    if var_probit_calib = ' X_wc_clog ' then do;
        mse_probit_calib=(est_probit_calib - b1)**2;
        output;
    end;
    keep var_probit_calib mse_probit_calib;
run;
proc univariate data=b_probit_calib;by var_probit_calib;
var mse_probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;

title 'Probit Model (Y_sim_probit with X*_wg)';
proc logistic data=Y_sim_probit;by rep;
    model Y (event='1')= X_wg_clog /link=normit;
    ods output parameterestimates=b_probit_calib;
run;

data b_probit_calib;set b_probit_calib;by rep;
    var_probit_calib=Variable;
    est_probit_calib=Estimate;
    keep rep var_probit_calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_probit_calib;set b_probit_calib;
b1 = 0.371;
    if var_probit_calib = ' X_wg_clog ' then do;
        mse_probit_calib=(est_probit_calib - b1)**2;
        output;
    end;

```

```

        end;
        keep var_probit_calib mse_ probit_calib;
run;
proc univariate data=b_probit_calib;by var_probit_calib;
var mse_probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;
title 'Cloglog Model (Y_sim_clog with X*_wc)';
proc logistic data=Y_sim_clog;by rep;
model Y (event='1')= X_wc_clog /link=cloglog;
ods output parameterestimates=b_clog_calib;
run;

data b_clog_calib;set b_clog_calib;by rep;
var_clog_calib=Variable;
est_clog_calib=Estimate;
keep rep var_clog_calib est_clog_calib;
run;
proc sort data=b_clog_calib;by var_clog_calib;
run;

data b_clog_calib;set b_clog_calib;
b1 = 0.371;
if var_clog_calib = ' X_wc_clog ' then do;
mse_clog_calib=(est_clog_calib - b1)**2;
output;
end;
keep var_clog_calib mse_clog_calib;
run;
proc univariate data=b_clog_calib;by var_clog_calib;
var mse_clog_calib;
output out=mse_clog_calib mean=mse_clog_calib;
run;

title 'Cloglog Model (Y_sim_clog with X*_wg)';
proc logistic data=Y_sim_clog;by rep;
model Y (event='1')= X_wg_clog /link=cloglog;
ods output parameterestimates=b_clog_calib;
run;

data b_clog_calib;set b_clog_calib;by rep;
var_clog_calib=Variable;
est_clog_calib=Estimate;
keep rep var_clog_calib est_clog_calib;
run;
proc sort data=b_clog_calib;by var_clog_calib;
run;

data b_clog_calib;set b_clog_calib;
b1 = 0.371;
if var_clog_calib = ' X_wg_clog ' then do;
mse_clog_calib=(est_clog_calib - b1)**2;
output;
end;
keep var_clog_calib mse_clog_calib;
run;
proc univariate data=b_clog_calib;by var_clog_calib;

```

```
var mse_ clog_calib;
output out=mse_clog _calib mean=mse_clog_calib;
run;
```

9. การสร้างโปรแกรมสำหรับการนำค่าพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ ตัวแปรอธิบาย X^*_wg จากเทคนิคการสอบเทียบลอจิสติก เพื่อประมาณค่าพารามิเตอร์และค่าความคลาดเคลื่อนเฉลี่ยกำลังสองภายใต้ตัวแบบ GLMs

```
title 'Logit Model (Y_sim_logit with X*_wc)';
proc logistic data=Y_sim_logit;by rep;
  model Y (event='1')= X_wc_logistic /link=logit;
  ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
  var_logit_calib=Variable;
  est_logit_calib=Estimate;
  keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
  if var_logit_calib = ' X_wc_logistic ' then do;
    mse_logit_calib=(est_logit_calib - b1)**2;
    output;
  end;
  keep var_logit_calib mse_logit_calib;
run;
proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Probit Model (Y_sim_probit with X*_wc)';
proc logistic data=Y_sim_probit;by rep;
  model Y (event='1')= X_wc_logistic /link=normit;
  ods output parameterestimates=b_probit_calib;
run;

data b_ probit_calib;set b_probit_calib;by rep;
  var_probit_calib=Variable;
  est_probit_calib=Estimate;
  keep rep var_probit_calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_ probit_calib;set b_probit_calib;
b1 = 0.371;
  if var_ probit_calib = ' X_wc_logistic ' then do;
```

```

        mse_probit_calib=(est_ probit_calib - b1)**2;
        output;
    end;
    keep var_probit_calib mse_probit_calib;
run;

proc univariate data=b_probit_calib;by var_probit_calib;
var mse_probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;

title 'Cloglog Model (Y_sim_clog with X*_wc)';
proc logistic data=Y_sim_clog;by rep;
    model Y (event='1')= X_wc_logistic /link=cloglog;
    ods output parameterestimates=b_ clog_calib;
run;

data b_ clog_calib;set b_ clog_calib;by rep;
    var_ clog_calib=Variable;
    est_ clog_calib=Estimate;
    keep rep var_ clog_calib est_ clog_calib;
run;
proc sort data=b_ clog_calib;by var_ clog_calib;run;

data b_ clog_calib;set b_ clog_calib;b1 = 0.371;
    if var_ clog_calib = ' X_wc_logistic ' then do;
        mse_ clog_calib=(est_ clog_calib - b1)**2;
        output;end;
    keep var_ clog_calib mse_ clog_calib;
run;
proc univariate data=b_ clog_calib;by var_ clog_calib;
var mse_ clog_calib;
output out=mse_ clog_calib mean=mse_ clog_calib;
run;

```

10. การสร้างโปรแกรมสำหรับการนำค่าพยากรณ์หรือประมาณกลุ่มจริงของตัวแปรอธิบาย X^*_wc และ ตัวแปรอธิบาย X^*_wg จากเทคนิคการสอบเทียบลอจิสต์ เพื่อประมาณค่าพารามิเตอร์และค่าความคลาดเคลื่อนเฉลี่ยกำลังสองภายใต้ตัวแบบ GLMs

```

title 'Logit Model (Y_sim_logit with X*_wg)';
proc logistic data=Y_sim_logit;by rep;
    model Y (event='1')= X_wg_logit /link=logit;
    ods output parameterestimates=b_logit_calib;
run;

data b_logit_calib;set b_logit_calib;by rep;
    var_logit_calib=Variable;
    est_logit_calib=Estimate;
    keep rep var_logit_calib est_logit_calib;
run;
proc sort data=b_logit_calib;by var_logit_calib;
run;

```

```

data b_logit_calib;set b_logit_calib;
b1 = 0.371;
    if var_logit_calib = ' X_wg_logit ' then do;
        mse_logit_calib=(est_logit_calib - b1)**2;
        output;
    end;
keep var_logit_calib mse_logit_calib;
run;
proc univariate data=b_logit_calib;by var_logit_calib;
var mse_logit_calib;
output out=mse_logit_calib mean=mse_logit_calib;
run;

title 'Probit Model (Y_sim_probit with X*_wg)';
proc logistic data=Y_sim_probit;by rep;
model Y (event='1')= X_wg_logit /link=normit;
ods output parameterestimates=b_probit_calib;
run;

data b_probit_calib;set b_probit_calib;by rep;
var_probit_calib=Variable;
est_probit_calib=Estimate;
keep rep var_probit_calib est_probit_calib;
run;
proc sort data=b_probit_calib;by var_probit_calib;
run;

data b_probit_calib;set b_probit_calib;
b1 = 0.371;
    if var_probit_calib = ' X_wg_logit ' then do;
        mse_probit_calib=(est_probit_calib - b1)**2;
        output;
    end;
keep var_probit_calib mse_probit_calib;
run;
proc univariate data=b_probit_calib;by var_probit_calib;
var mse_probit_calib;
output out=mse_probit_calib mean=mse_probit_calib;
run;

title 'Cloglog Model (Y_sim_clog with X*_wg)';
proc logistic data=Y_sim_clog;by rep;
model Y (event='1')= X_wg_logit /link=cloglog;
ods output parameterestimates=b_clog_calib;
run;

data b_clog_calib;set b_clog_calib;by rep;
var_clog_calib=Variable;
est_clog_calib=Estimate;
keep rep var_clog_calib est_clog_calib;
run;
proc sort data=b_clog_calib;by var_clog_calib;
run;

data b_clog_calib;set b_clog_calib;
b1 = 0.371;
    if var_clog_calib = ' X_wg_logit ' then do;

```

```
        mse_clog_calib=(est_clog_calib - b1)**2;  
        output;  
    end;  
    keep var_clog_calib mse_clog_calib;  
run;  
proc univariate data=b_clog_calib;by var_clog_calib;  
var mse_clog_calib;  
output out=mse_clog_calib mean=mse_ clog_calib;  
run;
```

11. การสร้างไฟล์สำหรับเก็บผลลัพธ์

```
filename logfile 'd:\out.log';  
filename lisfile 'd:\out.lis';  
  
proc printto log=logfile print= lisfile new;  
run;
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ประวัติผู้วิจัย

ชื่อ-สกุล	นางสาวปานจิต วัฒนสารัช
ที่อยู่	9/30 ม.รัตนธิเบศร์ ต.เสาธงหิน อ.บางใหญ่ จ.นนทบุรี 11140
ประวัติการศึกษา	
พ.ศ. 2549	สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาสถิติ มหาวิทยาลัยศิลปากร
พ.ศ. 2550	ศึกษาต่อระดับปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์ มหาวิทยาลัยศิลปากร

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์