อัตราส่วนที่เหมาะสมของการเกิดเหตุการณ์ที่สนใจต่อตัวแปรสำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาคในกรณีที่มีความคลาดเคลื่อนในตัวแปรอธิบาย

The Optimality of the Number of Events per Variable for Binary Logistic Regression with Misclassifed Covariates

ศิวพร ภูกองทอง*

สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี

ตำบลหมากแข้ง อำเภอเมือง จังหวัดอุดรธานี 41000

แสงหล้า ชัยมงคล

ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Siwaporn Phukongtong*

Department of Applied Statistics, Faculty of Science, Udon Thani Rajabhat University,

Mark Khaeng, Muang, Udon Thani 41000

Saengla Chaimonkol

Department of Biotechnology, Faculty of Science and Technology, Thammasat University,

Rangsit Centre, Khlong Nueng, Khlong Luang, Pathum Thani 12120

บทคัดย่อ

งานวิจัยนี้ผู้วิจัยศึกษาอัตราส่วนที่เหมาะสมของการเกิดเหตุการณ์ที่สนใจต่อตัวแปรอธิบาย (events per variable หรือ EPV) สำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค เมื่อมีความผิดพลาดจากการวัดหรือการจำแนกผิดในตัวแปรอธิบาย โดยใช้วิธีการสุ่มตัวอย่างข้อมูลของการคุมกำเนิดของหญิงไทยที่สมรสอายุ 15-59 ปี จำนวน 25,812 คนที่ได้จากการสำรวจภาวะเจริญพันธุ์ พ.ศ. 2552 โดยสำนักงานสถิติแห่งชาติ ที่มีตัวแปรที่สนใจคือการคุมกำเนิดของหญิงไทย (Y) ที่คุมกำเนิดคิดเป็นร้อยละ 72.8 โดยมีตัวแปรอธิบาย (X) 4 ตัว ได้แก่อายุของหญิงไทยที่สมรส จำนวนบุตรเกิดรอดที่มีชีวิต ศาสนา และความต้องการในการมีบุตรเพิ่ม ข้อมูลชุดนี้มี EPV เท่ากับ (18,795/4) เท่ากับ 4,698.75 ในการศึกษาครั้งนี้กำหนด EPV 4 ระดับ คือ 10, 15, 20 และ 25 ขนาดตัวอย่างเท่ากับ 1,000 อัตราจำแนกผิดในตัวแปรอธิบายเชิงกลุ่มแบบทวิภาค 3 ระดับ คือ 2, 5 และ 10 % ของขนาดตัวอย่างที่กำหนด และเทอมของความผิดพลาดที่เกิดจากการวัดในตัวแปรอธิบายแบบต่อเนื่องและไม่ต่อเนื่อง (e) ใช้

***ผู้รับผิดชอบบทความ :** siwaporn.pkt@gmail.com DOI 10.14456/tstj.2015.16

การจำลองข้อมูลโดยกำหนดให้ $e \sim N(0, \sigma_e^2)$ และกำหนด σ_e^2 เป็น 3 ระดับ คือ 1, 5 และ 10 การจำลองและการวิเคราะห์ใช้โปรแกรม R เวอร์ชัน 3.0.2 ทำซ้ำจำนวน 5,000 รอบ ในแต่ละสถานการณ์ และมีเกณฑ์ที่ใช้พิจารณาในการเลือก EPV ที่เหมาะสม ได้แก่ ค่าเอนเอียงสัมพัทธ์และอำนาจการทดสอบ โดยการเปรียบเทียบค่าที่ได้จากตัวแบบที่มีความผิดพลาดในตัวแปรอธิบายกับตัวแบบที่ไม่มีความผิดพลาดในตัวแปรอธิบาย ผลการศึกษาสรุปได้ดังนี้ ในกรณีที่มีความผิดพลาดจากการวัดหรือการจำแนกผิดในตัวแปรอธิบายเพียง 1 ตัวของตัวแบบการถดถอยลอจิสติกทวิภาค ค่า EPV ที่เหมาะสมไม่ควรต่ำกว่า 15 ซึ่งสูงกว่ากฎ rule of thumb ของค่า EPV ที่กำหนดให้ไม่ต่ำกว่า 10 นอกจากนี้ยังพบว่าประเภทและลักษณะของตัวแปรอธิบายที่มีความผิดพลาด และขนาดของความผิดพลาดที่เพิ่มขึ้นมีผลต่อค่า EPV ให้มีแนวโน้มมากขึ้น

คำสำคัญ : ตัวแบบเชิงเส้นนัยทั่วไป; ความผิดพลาดจากการวัด; การจำแนกผิด

Abstract

The objective of this research is to evaluate the optimality of the number of events per variable (EPV) for binary logistic regression when covariate is misclassified or measured with error. The simulation based on data from the 2009 Reproductive Health Survey by National Statistical Office, Thailand. The variable of interest is the contraception in married Thai women aged between 15 and 59 and there are 4 covariates. The number of events per explanatory variable was (18,795/4) 4,698.75 for the full data. We considered values of EPV = 10, 15, 20 and 25; sample sizes of 1,000; misclassification rates of 2, 5 and 10 % of sample sizes; and measurement error term is $e \sim N(0, \sigma_e^2)$, σ_e^2 = 1, 5 and 10. The 5,000 data sets were simulated and then analyzed in R program, version 3.0.2. Results of the model with covariate misclassification or measurement error were compared for relative bias and power of the test against the results of model without covariate misclassification or measurement error. We found that the optimal of EPV values should no less than 15 when covariate is misclassified or measured with error. Moreover, type of covariate misclassification or measurement error and size of error influence on EPV.

Keywords: generalized linear model; measurement error; misclassification error

1. บทนำ

ในปัจจุบันงานวิจัยทั้งทางด้านวิทยาศาสตร์ การแพทย์ ชีวสถิติ ธุรกิจ เศรษฐศาสตร์ และสังคมศาสตร์ เป็นต้น มีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรที่ต้องการศึกษาหรือตัวแปรตอบสนองกับตัวแปรที่เป็นปัจจัยต่าง ๆ ที่เกี่ยวข้องกับตัวแปรที่เราต้องการศึกษาหรือตัวแปรอธิบายที่เป็นได้ทั้งตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพ ดังนั้นผู้วิจัยจำเป็นที่ต้องเลือกตัวแบบสำหรับการวิเคราะห์ให้เหมาะสมดังวัตถุประสงค์และชนิดของข้อมูล ถ้าตัวแปรตอบสนองเป็นข้อมูลเชิงคุณภาพ เช่น ข้อมูลเชิงกลุ่มหรือข้อมูลนับ ตัวแบบที่นิยมและใช้ได้ดีในกรณีนี้ คือ ตัวแบบการ

ถดถอยลอจิสติก โดยทั่วไปการใช้การวิเคราะห์การถดถอยลอจิสติก ผู้วิจัยสมมติว่าตัวแปรอธิบายต่าง ๆ นั้นได้มาอย่างถูกต้อง อย่างไรก็ตามในสภาพความเป็นจริงการวัดหรือการบันทึกข้อมูลอาจจะเกิดความผิดพลาดขึ้นได้ เมื่อเกิดความผิดพลาดจากการวัดหรือการบันทึกข้อมูลในตัวแปรอธิบาย ความผิดพลาดนี้อาจเป็นตัวแปรสุ่มหรือเป็นระบบ (ไม่สุ่ม) และความผิดพลาดที่เกิดขึ้นในตัวแปรเชิงปริมาณ เรียกว่าความผิดพลาดจากการวัด (measurement error) ส่วนความผิดพลาดที่เกิดขึ้นในตัวแปรเชิงคุณภาพ เรียกว่าการจำแนกผิด (misclassification error) ซึ่งความผิดพลาดที่เกิดขึ้นนี้จะส่งผลกระทบต่อประสิทธิภาพการประมาณค่าและอำนาจการทดสอบสมมติฐานของพารามิเตอร์ ดังจะเห็นได้จากงานวิจัยของ Mote และ Anderson ที่พบว่าการจำแนกผิดในตัวแปรอธิบายส่งผลให้อำนาจการทดสอบลดลง [1] และ Rosner และคณะ ที่พบว่าในการวิเคราะห์การถดถอยลอจิสติกเมื่อตัวแปรอธิบายมีความผิดพลาดจากการวัดจะทำให้ค่าประมาณพารามิเตอร์ที่ได้เกิดความเอนเอียง [2] นอกจากนี้จำนวนเหตุการณ์ที่สนใจต่อตัวแปรอธิบายที่มีอิทธิพลต่อผลตอบสนอง หรือที่เรียกว่าอัตราส่วนของการเกิดเหตุการณ์ที่สนใจต่อตัวแปรอธิบาย (events per variable หรือ EPV) เป็นปัจจัยหนึ่งที่ส่งผลต่อประสิทธิภาพการประมาณค่าและอำนาจการทดสอบสมมติฐานของพารามิเตอร์ ดังจะเห็นได้จากงานวิจัยของ Katz ที่ได้ศึกษาเกี่ยวกับการสูบบุหรี่ในช่วงวัยรุ่นกับการเกิดภาวะตื่นตระหนกในวัยผู้ใหญ่ตอนต้นที่ใช้ตัวอย่างจำนวน 688 คน และมีตัวแปรอธิบายที่มีอิทธิพลต่อผลตอบสนองจำนวน 8 ตัว ซึ่งข้อมูลชุดนี้มีค่า EPV เท่ากับ 0.88 ซึ่งถือว่าน้อยมาก ๆ ทำให้มีผลต่อช่วงความเชื่อมั่นของอัตราส่วน odds โดยทำให้ช่วงความเชื่อมั่นกว้างขึ้น [3] และ Peduzzi และคณะ ได้เสนอกฎ rule of thumb ของ EPV ในการวิเคราะห์การถดถอยลอจิสติกว่าควรจะมีค่า EPV ที่ไม่น้อยกว่า 10 เพราะถ้าต่ำกว่านี้จะทำให้ค่าความเอนเอียงของค่าประมาณพารามิเตอร์เพิ่มมากขึ้นและส่งผลต่อการทดสอบสมมติฐานของพารามิเตอร์ [4] แต่ในปี ค.ศ. 2006 Vittinghoff และ McCulloch ได้พยายามที่จะลดค่า EPV ลง โดยทำการศึกษาด้วยการจำลองข้อมูลภายใต้ EPV ในช่วง 2-16 ซึ่งมีเกณฑ์ที่ใช้พิจารณาได้แก่ ความเอนเอียงสัมพัทธ์ ช่วงความเชื่อมั่นที่คลุมค่าจริง ความผิดพลาดประเภทที่ 1 และอำนาจการทดสอบของพารามิเตอร์ และได้ข้อสรุปว่ากฎ rule of thumb ของ EPV นี้สามารถลดลงได้แต่ไม่ควรมีค่าน้อยกว่า 5 [5] แต่อย่างไรก็ตาม กฎ rule of thumb ที่เสนอโดย Peduzzi และคณะก็เป็นที่ยอมรับโดยทั่วไป ดังนั้นในงานวิจัยนี้ผู้วิจัยจึงสนใจหาค่า EPV ที่เหมาะสม สำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค เมื่อมีความผิดพลาดในตัวแปรอธิบาย นอกจากนี้ผู้วิจัยยังสนใจศึกษาอิทธิพลของประเภทและลักษณะของตัวแปรอธิบายที่มีความผิดพลาดจากการวัดหรือการจำแนกผิด และขนาดของความผิดพลาดในตัวแปรอธิบายที่มีผลต่อค่า EPV ภายใต้ตัวแบบการถดถอยลอจิสติกทวิภาค ซึ่งใช้วิธีการสุ่มตัวอย่างจากข้อมูลการคุมกำเนิดของผู้หญิงไทยที่สมรสอายุ 15-59 ปี จำนวน 25,812 คนจากการสำรวจภาวะเจริญพันธุ์ พ.ศ. 2552 โดยสำนักงานสถิติแห่งชาติและกรมอนามัย

2. การออกแบบการจำลองข้อมูล

การศึกษาครั้งนี้ใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ โดยมีขนาดตัวอย่างจำนวน 1,000 คน จากข้อมูลการคุมกำเนิดของหญิงไทยที่สมรสและมีอายุอยู่ในช่วง 15-59 ปี ที่ได้จากการสำรวจภาวะเจริญพันธุ์ พ.ศ. 2552 โดยสำนักงานสถิติแห่งชาติและกรมอนามัย จำนวน 25,812 คน โดยมีตัวแปรที่สนใจ คือหญิงไทยมีการคุมกำเนิด (Y=1) และ (Y=0) กรณีอื่น ๆ ข้อมูลนี้

มี Y=1 จำนวน 18,795 หรือคิดเป็นร้อยละ 72.8 และตัวแปรอธิบายที่มีผลต่อการคุมกำเนิดมีจำนวน 4 ตัว ได้แก่ X_1 คืออายุของหญิงไทยที่สมรส X_2 คือจำนวนบุตรเกิดรอดที่มีชีวิต X_3 คือศาสนา (โดยที่ X_3=1 หมายถึงอิสลาม และ X_3=0 กรณีอื่น ๆ) และ X_4 คือความต้องการในการมีบุตรเพิ่ม (โดยที่ X_4=1 หมายถึงไม่ต้องการ และ X_4=0 หมายถึงต้องการ) สถิติพรรณนาของตัวแปรตอบสนองและตัวแปรอธิบายทั้ง 4 ตัว แสดงในตารางที่ 3 การศึกษาครั้งนี้ได้กำหนดให้มีจำนวนของตัวแปรอธิบาย (X) ที่มีความผิดพลาดในตัวแบบเพียง 1 ตัว ในกรณีการจำแนกผิดในตัวแปร X_3 จะกำหนดอัตราการจำแนกผิด 2 ระดับ คือ 2 และ 5 % เนื่องจากข้อมูลทั้งหมดมีสัดส่วนของ X_3=1 เพียงร้อยละ 6 และการจำแนกผิดในตัวแปร X_4 จะกำหนดอัตราการจำแนกผิด 3 ระดับ คือ 2, 5 และ 10 % ของขนาดตัวอย่างที่กำหนด โดยมีการจำแนกผิดใน 2 ลักษณะ คือ การจำแนกค่า 1 ผิดเป็น 0 และการจำแนกค่า 0 ผิดเป็น 1 ส่วนกรณีมีความผิดพลาดจากการวัดในตัวแปร X_1 หรือ X_2 จะกำหนดเทอมของความผิดพลาดที่เกิดจากการวัดในตัวแปร X_1 และ X_2 (e_i; i=1,2) โดยใช้การจำลองข้อมูล $e_i \sim N(0,\sigma^2_{e_i})$ ซึ่งกำหนด $\sigma^2_{e_i}$ เป็น 3 ระดับ คือ 1, 5 และ 10 ข้อมูลชุดนี้มี EPV คือ (18,795/4) เท่ากับ 4,698.75 แต่การศึกษาครั้งนี้กำหนดให้มี EPV 4 ระดับ คือ 10, 15, 20 และ 25 การจำลองและการวิเคราะห์ข้อมูลใช้โปรแกรม R เวอร์ชัน 3.0.2 ทำซ้ำจำนวน 5,000 รอบในแต่ละสถานการณ์

3. วิธีดำเนินการศึกษา

ทำการแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มของ Y=1 และ Y=0 แล้วสุ่มตัวอย่างจากแต่ละกลุ่มของ Y อย่างเป็นอิสระกัน โดยให้ได้ตามระดับของ EPV ที่กำหนด นั่นคือ จำนวนเหตุการณ์ที่สนใจ (Y=1) คำนวณมาจากนำ EPV คูณกับจำนวนตัวแปรอธิบายในตัวแบบ ดังนั้น ณ ระดับของ EPV เท่ากับ 10, 15, 20 และ 25 จะสุ่มตัวอย่างของกลุ่ม Y=1 จำนวน 40, 60, 80 และ 100 ตามลำดับ และสุ่มตัวอย่างจากกลุ่ม Y=0 เพื่อให้ได้ตามขนาดตัวอย่าง (n) ที่กำหนด แสดงดังตารางที่ 1

ตารางที่ 1 จำนวนของตัวแปร Y ที่ต้องสุ่มตามระดับของ EPV

ระดับของ EPV	จำนวนของ Y = 1	จำนวนของ Y = 0
10	40	n - 40
15	60	n - 60
20	80	n - 80
25	100	n - 100

จากนั้นจำลองข้อมูลของตัวแปรอธิบายที่มีความผิดพลาดจากการวัดหรือการจำแนกผิด (W_j) เมื่อ j=1,2,3,4,5,6 ดังต่อไปนี้ $W_1 = X_1 + e_1$ เมื่อ $e_1 \sim N(0,\sigma^2_{e_1})$, $W_2 = X_2 + e_2$ เมื่อ $e_2 \sim N(0,\sigma^2_{e_2})$, W_3 จำแนกผิดที่ได้จากการแทนค่า 1 ด้วย 0 ของค่าของตัวแปร X_3 , W_4 จำแนกผิดที่ได้จากการแทนค่า 0 ด้วย 1 ของค่าของตัวแปร X_3 , W_5 จำแนกผิดที่ได้จากการแทนค่า 1 ด้วย 0 ของค่าของตัวแปร X_4 และ W_6 จำแนกผิดที่ได้จากการแทนค่า 0 ด้วย 1 ของค่าของตัวแปร X_4 แล้วนำไปสร้างตัวแบบการถดถอยลอจิสติกทวิภาค ดังตารางที่ 2

จากนั้นทำการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดของฟิชเชอร์สกอริ่ง (Fisher's scoring method) แล้วคำนวณค่าความเอนเอียงสัมพัทธ์ (relative bias) ของค่าประมาณพารามิเตอร์ ($\hat{\beta}_j$) และอำนาจการทดสอบ (power of the test) ของพารามิเตอร์ในการทดสอบ $H_0: \beta_j = 0$ เมื่อ j=1, 2,3,4 โดยใช้ตัวสถิติอัตราส่วนภาวะน่าจะเป็น

(likelihood ratio test) โดยมีรายละเอียดในการคำนวณดังนี้

3.1 ความเอนเอียงสัมพัทธ์

$$\text{ความเอนเอียงสัมพัทธ์ของ } \hat{\beta}_j = \left| \frac{\hat{\beta}_{jm,Error} - \hat{\beta}_{jm,M0}}{\hat{\beta}_{jm,M0}} \right| \times 100\% \quad ; j=1,2,3,4 \; m=1,2,\ldots, 5000 \qquad (1)$$

เมื่อ $\hat{\beta}_{jm,M0}$ คือค่าประมาณพารามิเตอร์ตัวที่ j ของตัวแบบที่ไม่มีความผิดพลาดในตัวแปรอธิบาย (ตัวแบบ M0) ในการทำซ้ำรอบที่ m และ $\hat{\beta}_{jm,Error}$ คือค่าประมาณพารามิเตอร์ตัวที่ j ของตัวแบบที่มีความผิดพลาดจากการวัดหรือการจำแนกผิดในตัวแปรอธิบาย (ตัวแบบ M1, M2, M3, M4, M5, M6) ในการทำซ้ำรอบที่ m

3.2 อำนาจการทดสอบ

อำนาจการทดสอบของพารามิเตอร์ในการทดสอบ $H_0: \beta_j = 0$ เมื่อ j =1,2,3,4 โดยใช้ตัวสถิติอัตราส่วนภาวะน่าจะเป็น ซึ่งจะทำโดยการนับจำนวนครั้งที่ปฏิเสธ H_0 เมื่อ H_0 ไม่เป็นจริงและบวกสะสมค่าไว้ เมื่อทำครบทุกรอบแล้วก็จะนำค่าสะสมที่ได้มาหารด้วยจำนวนรอบ คือ 5,000 ครั้ง โดยจะพิจารณาค่าอำนาจการทดสอบที่ไม่ต่ำกว่าร้อยละ 80 เนื่องจากขนาดตัวอย่างที่ใช้มีขนาดใหญ่

ตารางที่ 2 รหัสตัวแบบและตัวแบบการถดถอยลอจิสติกทวิภาคที่ใช้ในการศึกษา

รหัสตัวแบบ	ตัวแบบการถดถอยลอจิสติกทวิภาคที่ใช้ในการศึกษา
M0	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
M1	$\text{logit}(\pi_i) = \beta_0 + \beta_1 w_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
M2	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 w_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
M3	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 w_{i3} + \beta_4 x_{i4}$
M4	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 w_{i4} + \beta_4 x_{i4}$
M5	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 w_{i5}$
M6	$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 w_{i6}$

ตารางที่ 3 สถิติพรรณนาของตัวแปรต่าง ๆ ของข้อมูลการคุมกำเนิดในหญิงไทยที่สมรส จำนวน 25,812 คน

ตัวแปร	ประเภทและลักษณะของตัวแปร	ค่าเฉลี่ย	ฐานนิยม	ร้อยละของการเกิดเหตุการณ์ที่สนใจ
อายุ (X_1)	ตัวแปรแบบต่อเนื่อง	39.511	-	-
จำนวนบุตรเกิดรอดที่มีชีวิต (X_2)	ตัวแปรแบบไม่ต่อเนื่อง	-	2	-
ศาสนา (X_3)	ตัวแปรเชิงกลุ่มแบบทวิภาค	-	-	0.063
ความต้องการในการมีบุตรเพิ่ม (X_4)	ตัวแปรเชิงกลุ่มแบบทวิภาค	-	-	0.832
การคุมกำเนิดในหญิงไทย (Y)	ตัวแปรเชิงกลุ่มแบบทวิภาค	-	-	0.728

4. ผลการวิจัย

4.1 สถิติพรรณนาของตัวแปรต่าง ๆ ของข้อมูล

ตารางที่ 3 แสดงสถิติต่าง ๆ ของข้อมูลทั้งหมด โดยมีตัวแปรที่สนใจ คือ หญิงไทยที่มีการคุมกำเนิด คิดเป็นร้อยละ 72.8 ตัวแปรอธิบายที่มีอิทธิพลต่อการคุมกำเนิดจำนวน 4 ตัว ได้แก่ อายุ จำนวนบุตรเกิดรอดที่มีชีวิต ศาสนา และความต้องการในการมีบุตรเพิ่ม ซึ่งในข้อมูลชุดนี้หญิงไทยมีอายุโดยเฉลี่ย 39.51 ปี มีจำนวนบุตรเกิดรอดที่มีชีวิตส่วนใหญ่ 2 คน และมีคนที่ต้องการมีบุตรเพิ่ม (X_4=1) คิดเป็นร้อยละ 83 และเป็นผู้ที่นับถือศาสนาอิสลาม (X_3=1) คิดเป็นร้อยละ 6 เมื่อนำเอาข้อมูลทั้งหมดนี้ไปสร้างตัวแบบการถดถอยลอจิสติกทวิภาค มีรูปแบบดังนี้

$$\mathrm{logit}(\pi) = 3.080 - 0.104x_1 + 0.380x_2 - 1.570x_3 + 1.938x_4 \quad (2)$$

ตารางที่ 4 ค่าเอนเอียงสัมพัทธ์ของค่าประมาณพารามิเตอร์และอำนาจการทดสอบของพารามิเตอร์ของตัวแปรอธิบายที่มีความผิดพลาดในตัวแบบต่าง ๆ

ตัวแบบ	เกณฑ์เปรียบเทียบ	ขนาดของความผิดพลาด	EPV			
			10	15	20	25
M1	Relative bias	$\sigma^2_{e_1}$ = 1	1.538	1.297	1.251	1.190
		$\sigma^2_{e_1}$ = 5	18.821	18.993	19.508	19.482
		$\sigma^2_{e_1}$ = 10	49.473	49.616	50.187	50.376
	Power of the test	$\sigma^2_{e_1}$ = 1	0.781	0.925	0.976	0.992
		$\sigma^2_{e_1}$ = 5	0.704	0.872	0.949	0.977
		$\sigma^2_{e_1}$ = 10	0.512	0.700	0.823	0.890
M2	Relative bias	$\sigma^2_{e_1}$ = 1	40.179	39.963	40.298	41.050
		$\sigma^2_{e_1}$ = 5	94.718	94.550	95.048	94.900
		$\sigma^2_{e_1}$ = 10	98.711	98.652	98.603	98.625
M2	Power of the test	$\sigma^2_{e_1}$ = 1	0.446	0.632	0.764	0.853
		$\sigma^2_{e_1}$ = 5	0.090	0.108	0.115	0.138
		$\sigma^2_{e_1}$ = 10	0.061	0.065	0.071	0.071
M3	Relative bias	2 %	18.322	18.672	18.955	19.416
		5 %	48.639	50.013	51.934	53.002
	Power of the test	2 %	0.644	0.808	0.895	0.942
		5 %	0.279	0.366	0.436	0.489
M4	Relative bias	2 %	6.381	6.744	7.040	7.105
		5 %	15.207	16.155	16.802	17.533
	Power of the test	2 %	0.916	0.980	0.995	0.998
		5 %	0.972	0.996	0.999	1.000

ตารางที่ 4 (ต่อ)

ตัวแบบ	เกณฑ์เปรียบเทียบ	ขนาดของความผิดพลาด	EPV			
			10	15	20	25
M5	Relative bias	2 %	4.160	4.256	4.377	4.520
		5 %	10.319	10.665	11.075	11.249
		10 %	19.511	19.993	20.638	20.940
	Power of the test	2 %	0.993	0.999	1.000	1.000
		5 %	0.998	1.000	1.000	1.000
		10 %	1.000	1.000	1.000	1.000
M6	Relative bias	2 %	11.457	11.763	12.283	12.524
		5 %	27.821	28.511	29.478	30.159
		10 %	49.392	50.850	52.420	53.502
	Power of the test	2 %	0.951	0.991	0.998	1.000
		5 %	0.811	0.932	0.972	0.990
		10 %	0.464	0.602	0.689	0.758

นอกจากนี้ยังพบว่าตัวแปรอธิบายที่มีความสัมพันธ์กันอย่างชัดเจน คือ (1) X_1 มีความสัมพันธ์กับ X_2 และ X_4 (2) X_2 มีความสัมพันธ์กับ X_4

4.2 ผลการศึกษาของอัตราส่วนของการเกิดเหตุการณ์ที่สนใจต่อตัวแปรอธิบาย (EPV) สำหรับตัวแบบการถดถอยลอจิสติกทวิภาค ในกรณีที่มีความผิดพลาดในตัวแปรอธิบาย

จากตารางที่ 4 เมื่อพิจารณาจากค่าเอนเอียงสัมพัทธ์ของค่าประมาณพารามิเตอร์ของตัวแปรที่มีการวัดผิดพลาด พบว่าค่าเอนเอียงสัมพัทธ์มีแนวโน้มไม่แตกต่างกันในแต่ละระดับของ EPV ณ ระดับหนึ่ง ๆ ของขนาดของความผิดพลาด แต่แปรผันตรงกับขนาดของความผิดพลาดในทุกตัวแบบ และเมื่อพิจารณาจากอำนาจการทดสอบพารามิเตอร์ของตัวแปรที่มีการวัดผิดพลาด พบว่าอำนาจการทดสอบแปรผันตรงกับระดับของ EPV แต่แปรผกผันกับขนาดของความผิดพลาดในทุกตัวแบบ ยกเว้น M4 และ M5 ที่อำนาจการทดสอบแปรผันตรงกับระดับของ EPV และขนาดของความผิดพลาด

เมื่อพิจารณาจากทั้งค่าเอนเอียงสัมพัทธ์ของค่าประมาณพารามิเตอร์และอำนาจการทดสอบพารามิเตอร์ของตัวแปรอธิบายที่มีการวัดผิดพลาด พบว่าค่า EPV ที่เหมาะสมไม่ควรต่ำกว่า 15 สำหรับตัวแปรต่อเนื่องที่มีการวัดผิดพลาด และค่า EPV ไม่ควรต่ำกว่า 10 สำหรับตัวแปรเชิงกลุ่มแบบทวิภาคที่มีการจำแนกผิด นอกจากนี้ยังพบว่าผลของความผิดพลาดที่เกิดขึ้นในตัวแปรอธิบายที่มีความผิดพลาด ส่งผลกระทบต่อตัวแปรอธิบายอื่นที่มีความสัมพันธ์กัน เช่น ตัวแปร X2 ได้รับผลกระทบจากการวัดผิดพลาดในตัวแปร X1 ทำให้ค่าเอนเอียงสัมพันธ์ของตัวแปร X2 มีค่าน้อยกว่าค่าเอนเอียงสัมพันธ์ของตัวแปร X1 เพียงเล็กน้อย และทำให้อำนาจการทดสอบของตัวแปร X2 มีค่าค่อนข้างต่ำโดยเฉพาะที่ระดับของ เท่ากับ 10 ดังแสดงในตารางที่ 5

ตารางที่ 5 ค่าเอนเอียงสัมพัทธ์ของค่าประมาณพารามิเตอร์และอำนาจการทดสอบของพารามิเตอร์ของตัวแปรอธิบาย X_1 และ X_2 ในตัวแบบ M1

ตัวแปร	เกณฑ์เปรียบเทียบ	ขนาดของความผิดพลาด	EPV			
			10	15	20	25
X_1	Relative bias	$\sigma^2_{e_1} = 1$	1.538	1.297	1.251	1.190
		$\sigma^2_{e_1} = 5$	18.821	18.993	19.508	19.482
		$\sigma^2_{e_1} = 10$	49.473	49.616	50.187	50.376
	Power of the test	$\sigma^2_{e_1} = 1$	0.781	0.925	0.976	0.992
		$\sigma^2_{e_1} = 5$	0.704	0.872	0.949	0.977
		$\sigma^2_{e_1} = 10$	0.512	0.700	0.823	0.890
X_2	Relative bias	$\sigma^2_{e_1} = 1$	1.646	1.445	1.349	1.262
		$\sigma^2_{e_1} = 5$	18.439	18.521	19.064	18.812
		$\sigma^2_{e_1} = 10$	46.279	47.005	46.611	46.493
	Power of the test	$\sigma^2_{e_1} = 1$	0.676	0.877	0.961	0.988
		$\sigma^2_{e_1} = 5$	0.545	0.777	0.890	0.957
		$\sigma^2_{e_1} = 10$	0.335	0.518	0.681	0.792

อย่างไรก็ตาม เมื่อพิจารณาในภาพรวมจากทั้งค่าเอนเอียงสัมพัทธ์ของค่าประมาณพารามิเตอร์และอำนาจการทดสอบพารามิเตอร์ของตัวแปรอธิบายทุกตัวในแต่ละตัวแบบ พบว่าค่า EPV ที่เหมาะสมไม่ควรต่ำกว่า 15 เกือบทุกกรณีที่ศึกษา ดังแสดงในตารางที่ 6

5. สรุป

งานวิจัยนี้เป็นการศึกษาเพื่อหาอัตราส่วนที่เหมาะสมของการเกิดเหตุการณ์ที่สนใจต่อตัวแปรอธิบาย (EPV) สำหรับการวิเคราะห์การถดถอยลอจิสติก ซึ่งสามารถนำไปประยุกต์ใช้กับงานวิจัยด้านต่าง ๆ ที่ใช้ข้อมูลจริงในการวิเคราะห์การถดถอยลอจิสติก ในกรณีที่ผู้วิจัยไม่แน่ใจในความถูกต้องจากการวัดหรือจำแนกผิดในตัวแปรอธิบาย และผู้วิจัยทำการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดของฟิชเชอร์สกอริ่ง (Fisher’s scoring method) และทดสอบความมีนัยสำคัญของพารามิเตอร์โดยใช้สถิติอัตราส่วนภาวะน่าจะเป็น (likelihood ratio test) สำหรับงานวิจัยนี้ผู้วิจัยทำการศึกษาค่า EPV ในช่วง 10 ถึง 25 และศึกษาเฉพาะกรณีที่มีความผิดพลาดในตัวแปรอธิบายเพียง 1 ตัว ในตัวแบบ ผลการศึกษาพบว่าในกรณีที่มีความผิดพลาดจากการวัดหรือการจำแนกผิดในตัวแปรอธิบาย 1 ตัว ค่า EPV ที่เหมาะสมไม่ควรต่ำกว่า 15 (ซึ่งสูงกว่ากฎ rule of thumb ของค่า EPV ที่ว่าควรมีค่า EPV ไม่ต่ำกว่า 10) เกือบทุกกรณีที่ศึกษา นอกจากนี้ยังพบว่าประเภทและลักษณะของตัวแปรอธิบายที่มีความผิดพลาด และขนาดของความผิดพลาดมีอิทธิพลต่อค่า EPV โดยพิจารณาจากแนวโน้ม เมื่อความผิดพลาดเพิ่มขึ้น ค่า EPV มีแนวโน้มมากขึ้น

ตารางที่ 6 สรุปผลการศึกษาของค่า EPV ที่เหมาะสมสำหรับตัวแบบการถดถอยลอจิสติกทวิภาคในกรณีที่มีความผิดพลาดในตัวแปรอธิบาย 1 ตัว

ตัวแปรที่มีการวัดผิดพลาด	ประเภทและลักษณะของตัวแปร	เทอมความผิดพลาด	อัตราการวัดผิดพลาด		
			$\sigma^2_{e_i} = 1$	$\sigma^2_{e_i} = 5$	$\sigma^2_{e_i} = 10$
อายุ (X_1)	ตัวแปรแบบต่อเนื่อง	$e_1 \sim N(0, \sigma^2_{e_1})$	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 25
จำนวนบุตรเกิดรอดที่มีชีวิต (X_2)	ตัวแปรแบบไม่ต่อเนื่อง	$e_2 \sim N(0, \sigma^2_{e_2})$	เท่ากับ 25	สรุปไม่ได้	สรุปไม่ได้
ตัวแปรที่มีการจำแนกผิด	**ประเภทและลักษณะของตัวแปร**	**ลักษณะการจำแนกผิด**	**อัตราการจำแนกผิด**		
			2 %	5 %	10 %
ศาสนา (X_3)	ตัวแปรเชิงกลุ่มแบบทวิภาค (สัดส่วน X_3=1 เท่ากับ 6 %)	จำแนกค่า 1 ผิดเป็น 0	ไม่ต่ำกว่า 15	สรุปไม่ได้	-
		จำแนกค่า 0 ผิดเป็น 1	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 15	-
ความต้องการในการมีบุตรเพิ่ม (X_4)	ตัวแปรเชิงกลุ่มแบบทวิภาค (สัดส่วน X_4=1 เท่ากับ 83 %)	จำแนกค่า 1 ผิดเป็น 0	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 15
		จำแนกค่า 0 ผิดเป็น 1	ไม่ต่ำกว่า 15	ไม่ต่ำกว่า 15	สรุปไม่ได้

6. ข้อเสนอแนะ

ในการวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะกรณีที่มีความผิดพลาดในตัวแปรอธิบายเพียงตัวเดียว ยังมีรูปแบบความผิดพลาดในตัวแปรลักษณะอื่น ๆ ได้แก่ กรณีที่มีความผิดพลาดในตัวแปรอธิบายมากกว่า 1 ตัว หรือมีความผิดพลาดในตัวแปรตอบสนอง เป็นต้น ดังนั้นในการวิจัยต่อไปอาจจะพิจารณาศึกษากรณีความผิดพลาดในตัวแปรลักษณะเหล่านี้ด้วย และข้อมูลที่ใช้ในการวิจัยครั้งนี้มีสัดส่วนของจำนวนเหตุการณ์ที่สนใจสูง คิดเป็นร้อยละ 72.8 ในการวิจัยต่อไปอาจศึกษาโดยใช้ข้อมูลที่มีสัดส่วนของจำนวนเหตุการณ์ที่สนใจในลักษณะอื่น ๆ อีกทั้งตัวแปรเชิงกลุ่มแบบทวิภาคในการวิจัยครั้งนี้มีสัดส่วนของการเกิดเหตุการณ์ที่สนใจน้อยมากและสูงมาก ในการวิจัยต่อไปอาจพิจารณาศึกษาในลักษณะอื่น ๆ และศึกษาค่า EPV ในช่วงที่มากกว่า 25

7. รายการอ้างอิง

[1] Mote, V.L. and Anderson, R.L., 1965, An investigation of the effect of misclassify-cation on the properties of χ^2 - tests in the analysis of categorical data, Biometrika 52: 95-109.

[2] Rosner, B., Spiegelman, D. and Willett, W.C., 1990, Correction of logistic regression relative risk estimates and confidence intervals for measurement

error: the case of multiple covariates measured with error, Am. J. Epidemiol. 132: 734-745.

[3] Katz, M.H., 2003, Multivariable analysis: A primer for readers of medical research, Ann. Intern. Med. 138: 644-650.

[4] Peduzzi, P., Concato, J., Kember, E., Holford, T.R. and Feinstein, A.R., 1996, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49: 1373-1379.

[5] Vittinghoff, E. and McCulloch, C.E., 2006, Relaxing the rule of ten events per variable in logistic and cox regression, Am. J. Epidemiol. 165: 710-718.