ตัวแบบการถดถอยที่มีผลกระทบจากค่าศูนย์ ประยุกต์ใช้กับจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ

ZERO-INFLATED REGRESSION MODEL APPLIED FOR VOLUNTARY MOTOR CLAIM INSURANCE

ปวริศา สุขเรื่อย[*] *สำรวม จงเจริญ*

Pawarisa Sukruey[*]***,*** *Samruam Chongcharoen*

คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

Faculty of Applied Statistics, National Institute of Development Administration.

***Corresponding author, e-mail:** s.pawarisa@hotmail.com*

บทคัดย่อ

ในธุรกิจประกันภัยรถยนต์นั้น การศึกษาปัจจัยที่มีผลต่อจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนสามารถประมาณการการเกิดความถี่ของความเสียหายได้ ซึ่งจะนำไปสู่การคำนวณต้นทุนความเสียหายได้อย่างแม่นยำขึ้น เพราะการคำนวณต้นทุนความเสียหายที่ถูกต้องและแม่นยำจะทำให้บริษัทสามารถกำหนดเบี้ยประกันภัยได้อย่างเพียงพอต่อค่าใช้จ่ายต่างๆ ที่จะเกิดขึ้นโดยเฉพาะการตั้งเงินสำรองจ่ายค่าสินไหมทดแทนได้อย่างเหมาะสม และสามารถจัดสรรเงินไปลงทุนต่อในแหล่งการลงทุนต่างๆ ได้อย่างคุ้มค่า โดยลักษณะข้อมูลของจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนมีค่าที่เป็นศูนย์อยู่มากอันเนื่องมาจากรถยนต์ที่อยู่ภายใต้เงื่อนไขกรมธรรม์นั้นไม่ได้เกิดอุบัติเหตุหรือผู้เอาประกันต้องรับผิดค่าเสียหายในส่วนแรกเอง ด้วยเหตุนี้งานวิจัยนี้จึงได้มีการประยุกต์การแจกแจงที่มีผลกระทบจากค่าศูนย์ เพื่อนำมาแก้ไขปัญหาข้อมูลที่ได้รับผลกระทบจากค่าศูนย์ โดยตัวแบบการถดถอยที่ใช้ในงานวิจัยนี้คือ ตัวแบบการถดถอยปัวซงที่มีผลกระทบจากค่าศูนย์ และตัวแบบการถดถอยทวินามนิเสธที่มีผลกระทบจากค่าศูนย์

จากการศึกษาพบว่า ตัวแบบการถดถอยที่ถูกเลือกจากการศึกษาครั้งนี้คือตัวแบบการถดถอยทวินามนิเสธที่มีผลกระทบจากค่าศูนย์ และผลการศึกษาปัจจัยที่มีผลต่อการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ คือ อายุของผู้เอาประกัน อายุของรถยนต์ ขนาดตัวถังรถยนต์ และรุ่นของรถยนต์

คำสำคัญ: จำนวนการเรียกร้องค่าสินไหม ตัวแบบการถดถอยที่มีผลกระทบจากค่าศูนย์ ข้อมูลที่มีค่าศูนย์อยู่มาก

Abstract

In the motor insurance business, the study of factors affecting the number of claims to estimate the frequency of damage and it can lead to calculate the loss cost precisely. The calculation of the loss cost that accuracy and precision will enable the company to set up adequate premium for the reserve and allocate money to invest in the resources to invest in cost-effectively. The appearance of the number of claims have an excess of zero counts since the car that are under the policy is not an accident or the insured has the deductible, fixed amount of an insurance claim that is the responsibility of the insured, and which the insurance company will deduct from the claim payment. For this reason, this research has been applied the zero-inflated regression model to solve problems of an excess of zero counts. The regression model that used in this research are zero-inflated Poisson regression model and zero-inflated negative binomial regression model.

The results found that the regression model selected from this study was the zero-inflated negative binomial regression model and the factors affecting the claim are the insured age, vehicle age, vehicle size, and model.

Keywords: Number of Claim, Zero-Inflated Regression Model, Excess Zero

บทนำ

ในธุรกิจประกันภัย การคำนวณต้นทุนความเสียหายนั้นต้องพิจารณาจากองค์ประกอบสองส่วนคือ ความถี่และความรุนแรงของการเกิดความเสียหาย ในส่วนของความถี่ของการเกิดความเสียหายหากบริษัทสามารถประมาณการณ์และรับทราบถึงปัจจัยที่มีผลต่อจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนของบริษัทประกันภัยได้ บริษัทก็จะสามารถกำหนดเบี้ยประกันภัยได้อย่างเพียงพอต่อค่าใช้จ่ายที่จะเกิดขึ้น โดยเฉพาะการตั้งเงินสำรองจ่ายค่าสินไหมทดแทนได้อย่างเหมาะสม อีกทั้งบริษัทยังสามารถจัดสรรเงินที่เก็บมาจากเบี้ยประกันภัยนอกเหนือจากการตั้งเงินสำรอง โดยการนำเบี้ยประกันภัยส่วนที่เหลือไปลงทุนต่อในแหล่งการลงทุนต่างๆ ได้อย่างคุ้มค่า

ธุรกิจประกันภัยที่กำลังเป็นที่นิยมในขณะนี้คือ ธุรกิจประกันภัยรถยนต์ เห็นได้จากบริษัทประกันวินาศภัยที่กำลังดำเนินธุรกิจในประเทศไทยต่างมีการรับประกันภัยรถยนต์ด้วยกันทั้งนั้น เพราะรถยนต์ได้เข้ามามีบทบาทสำคัญต่อการดำเนินชีวิตมากขึ้น เมื่อรถยนต์มีจำนวนมากขึ้นทำให้เกิดความหนาแน่นของการจราจร ส่งผลให้เกิดอุบัติเหตุบ่อยครั้ง เกิดความเสียหายไปยังชีวิตและทรัพย์สิน บริษัทจึงหาวิธีในการรองรับความเสียหายที่อาจจะเกิดขึ้น นั่นคือการทำประกันภัยรถยนต์นั่นเอง โดยลักษณะข้อมูลของจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนของประกันภัยรถยนต์นั้นมักพบข้อมูลที่มีค่าเป็นศูนย์อยู่มาก ทั้งนี้เนื่องจากรถยนต์ที่อยู่ภายใต้เงื่อนไขกรมธรรม์นั้นไม่ได้เกิดอุบัติเหตุ หรือหากเกิดอุบัติเหตุที่มีความเสียหายเพียงเล็กน้อย ผู้เอาประกันก็มักจะไม่แจ้งเพื่อเรียกร้องค่าสินไหมทดแทน อีกทั้งตามเงื่อนไขของประกันภัยรถยนต์มักจะพบเงื่อนไขการรับผิดค่าเสียหายส่วนแรก กล่าวคือผู้เอาประกันต้องรับภาระค่าใช้จ่ายส่วนแรกนี้เองหากเกิดอุบัติเหตุที่มีค่าความเสียหายไม่เกินจำนวนเงินตามที่

กรมธรรม์ได้ระบุไว้ หรือผู้เอาประกันไม่สามารถปฏิบัติตามเงื่อนไขที่กรมธรรม์ได้ระบุไว้ ดังนั้นจึงไม่เกิดการเรียกร้องค่าสินไหมทดแทนแก่กรมธรรม์นั้น ลักษณะข้อมูลของจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนจึงมีค่าที่เป็นศูนย์อยู่มาก

ด้วยเหตุนี้ ผู้วิจัยจึงเกิดความสนใจในการศึกษาหาปัจจัยที่มีผลต่อจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนของประกันภัยรถยนต์ภาคสมัครใจ และเมื่อพบว่าลักษณะข้อมูลของจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนนั้นมักพบข้อมูลที่เป็นค่าศูนย์อยู่มาก ผู้วิจัยจึงได้เลือกตัวแบบที่ต้องการศึกษาคือ ตัวแบบการถดถอยปัวซงที่มีผลกระทบจากค่าศูนย์ (Zero-Inflated Poisson Regression Model: ZIP) และตัวแบบการถดถอยทวินามนิเสธที่มีผลกระทบจากค่าศูนย์ (Zero-Inflated Negative Binomial Regression Model: ZINB)

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาปัจจัยที่มีผลต่อจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนที่มีค่าศูนย์อยู่จำนวนมาก โดยใช้ตัวแบบการถดถอยปัวซงที่มีผลกระทบจากค่าศูนย์ และตัวแบบการถดถอยทวินามนิเสธที่มีผลกระทบจากค่าศูนย์

2. เพื่อเปรียบเทียบความเหมาะสมของตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB โดยเกณฑ์ที่ใช้ในการคัดเลือกตัวแบบคือ เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's Information Criterion : AIC) และเกณฑ์สารสนเทศของเบส์ (Bayesian Information Criterion : BIC) ซึ่งการคัดเลือกตัวแบบการถดถอยที่ดีที่สุดพิจารณาจากค่า AIC และ BIC ที่มีค่าต่ำที่สุด

วิธีดำเนินการวิจัย

1. ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลทุติยภูมิการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ จากสมาคมวินาศภัยไทย โดยเป็นข้อมูลจากทุกบริษัทที่มีการทำประกันภัยรถยนต์ภาคสมัครใจในปีกรมธรรม์ 2557

2. ข้อมูลที่ได้ศึกษาจากงานวิจัยของ Karen C.H. Yip และ Kelvin K.W. Yau ในปี ค.ศ. 2005 [1] และงานวิจัยของ Noriszura Ismail และ Hossein Zamani ในปี ค.ศ. 2013 [2] พบว่า ข้อมูลที่จะใช้ในงานวิจัยครั้งนี้จะประกอบด้วย

2.1 ข้อมูลที่ใช้เป็นตัวแปรอธิบาย
- เพศของผู้ขับขี่ (Gender)
- อายุของผู้ขับขี่ (Age)
- อายุของรถยนต์ (Vehicle Age)
- ขนาดตัวถังของรถยนต์ (Vehicle Size)
- ยี่ห้อของรถยนต์ (Make)
- รุ่นของรถยนต์ (Model)

2.2 ข้อมูลที่ใช้เป็นตัวแปรตอบสนอง
- จำนวนครั้งในการเรียกร้องค่าสินไหมทดแทน (Number of Claim)

3. วิธีการวิเคราะห์ข้อมูล

3.1 ทำการตรวจสอบความถูกต้องของข้อมูล ตัดข้อมูลที่ไม่ครบถ้วนออก เช่น ข้อมูลที่ระบุเพศของผู้เอาประกัน อายุของผู้เอาประกัน อายุรถยนต์ ขนาดตัวถังรถยนต์ ยี่ห้อของรถยนต์ และรุ่นของรถยนต์ ที่ไม่ครบถ้วน

3.2 เนื่องจากรุ่นของรถยนต์มีจำนวนมาก จึงต้องมีการแบ่งกลุ่ม ตามการแบ่งกลุ่มรถยนต์นั่งสำหรับการประกันภัยรถยนต์ภาคสมัครใจ Andrew Leung [3] พบว่าในการวิจัยครั้งนั้น ผู้วิจัยใช้เกณฑ์ในการแบ่งคือ รถยนต์รุ่นใดที่มีจำนวนหน่วยเสี่ยงภัยมากกว่า 1% ของจำนวนรถยนต์ที่ทำประกันภัยทั้งหมด จะแยกรถยนต์รุ่นนั้นออกมา

พิจารณาต่างหาก ดังนั้น งานวิจัยครั้งนี้จึงใช้ 1% ของจำนวนรถยนต์ที่ทำประกันภัยทั้งหมดเช่นกัน โดยรถยนต์ที่ทำประกันภัยทั้งหมดในงานวิจัยชิ้นนี้คือ 21,445 คัน โดย 1% ของ 21,445 คือ 214.45 คัน ทางผู้วิจัยจึงปรับให้เป็นจำนวนเต็ม 200 คัน เพื่อความสะดวกในการนำไปใช้ในขั้นตอนต่อไป

โดยในงานวิจัยชิ้นนี้มีเกณฑ์การแบ่งกลุ่ม ดังนี้

1) รถยนต์รุ่นใดที่มีจำนวนหน่วยเสี่ยงภัย (จำนวนคันของรถยนต์) มากกว่า 200 หน่วย ภายใต้ยี่ห้อเดียวกัน จะแยกรถยนต์รุ่นนั้นออกมาพิจารณาต่างหาก

2) รถยนต์ต่างรุ่นที่มีจำนวนหน่วยเสี่ยงภัยน้อยกว่า 200 หน่วย แต่อยู่ภายใต้ยี่ห้อเดียวกัน จะถูกรวมเข้าด้วยกัน

3) เมื่อรวมรถยนต์ต่างรุ่นแต่อยู่ภายใต้ยี่ห้อเดียวกันแล้ว หากมีจำนวนหน่วยเสี่ยงภัยมากกว่า 200 หน่วย รถยนต์ยี่ห้อนั้นจะถูกนำมาพิจารณาต่างหาก แต่ถ้ามีจำนวนหน่วยเสี่ยงภัยน้อยกว่า 200 หน่วย จะถูกรวมเข้าด้วยกันในกลุ่มที่ชื่อว่า “อื่นๆ”

3.3 พิจารณาลักษณะข้อมูลเบื้องต้นของตัวแปรอธิบายที่คาดว่าจะส่งผลต่อจำนวนครั้งในการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ โดยใช้สถิติพรรณนา (Descriptive Statistics) ได้แก่ การคำนวณค่าทางสถิติพื้นฐาน เช่น ค่าเฉลี่ย (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ค่าความแปรปรวน (Variance) ค่าพิสัย (Range) และสร้างแผนภาพตรวจสอบลักษณะการกระจายของข้อมูล (Exploratory Data Analysis: EDA)

3.4 วิเคราะห์ข้อมูลเพื่อหาปัจจัยที่มีผลต่อการเรียกร้องค่าสินไหมทดแทน และสร้างสมการพยากรณ์จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทน โดยข้อมูลจำนวนครั้งในการเรียกร้องค่าสินไหมทดแทนเป็นค่าศูนย์อยู่จำนวนมาก ซึ่งใช้ตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB ในการวิเคราะห์ข้อมูล โดยมีรูปแบบของตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB โดยประมวลผลจากโปรแกรมสำเร็จ Statistical Analysis System (SAS) เวอร์ชัน 9.4 คำสั่ง Proc Genmod ซึ่งการวิเคราะห์ตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB มีรายละเอียดของตัวแบบ ดังนี้

1) ตัวแบบการถดถอยปัวซงที่มีผลกระทบจากค่าศูนย์ (Zero-inflated Poisson Regression Model : ตัวแบบการถดถอย ZIP) Lambert [4]

กำหนดให้ ตัวแปรสุ่ม $Y=(y_1,y_2,...,y_n)'$ เป็นเวกเตอร์ของตัวแปรตอบสนองที่มีขนาดเท่ากับ n ที่แต่ละ $y_i;i=1,2,3,...,n$ เป็นอิสระต่อกัน เป็นตัวแปรสุ่มของตัวแปรตอบสนองที่มีการแจกแจงแบบ ZIP มีพารามิเตอร์ 2 ตัว คือ พารามิเตอร์ค่าเฉลี่ย $\left(\underline{\lambda}=(\lambda_1,\lambda_2,...,\lambda_n)'\right)$, $\lambda_i>0$ ซึ่งเป็นเวกเตอร์ของค่าคาดหวังของตัวแปรตอบสนองที่มีขนาด $1\times n$ และ พารามิเตอร์ของความน่าจะเป็นที่จะเกิดศูนย์ $\left(\underline{p}=(p_1,p_2,...,p_n)'\right)$, $0\le p_i<1$, ซึ่งเป็นเวกเตอร์ของความน่าจะเป็นของสถานการณ์ที่จะเกิดศูนย์ที่มีขนาด เขียนการแจกแจงของตัวแปรสุ่มแทนได้ด้วย $Y\sim \text{ZIP}(\underline{\lambda},\underline{p})$ ซึ่ง

$$\begin{aligned} y_i&=0 && \text{ด้วยความน่าจะเป็น } p_i, i=1,2,...,n \\ y_i&\neq 0 && \text{ด้วยความน่าจะเป็น } 1-p_i, i=1,2,...,n \end{aligned} \qquad \text{.....1}$$

เพื่อทำให้ฟังก์ชันมวลความน่าจะเป็นของ เป็น

$$\Pr(Y = y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & ; y_i = 0, i = 1,2,3,....,n \\ (1-p_i)\left(\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) & ; y_i \neq 0, i = 1,2,3,..,n. \end{cases} \quad .....2$$

ที่มี $\underline{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)'$ และ $\underline{p} = (p_1, p_2, ..., p_n)'$ ที่สอดคล้องกับ

$\log(\lambda) = \underline{X}\beta$ เมื่อต้องการประมาณค่าของตัวแปรตอบสนอง

และ $\text{logit}(p) = \log\left(\dfrac{p}{1-p}\right) = \underline{G}\gamma$ เมื่อต้องการประมาณค่าความน่าจะเป็นที่จะเกิดศูนย์

เมื่อ X คือ เมทริกซ์ของตัวแปรอธิบายที่คาดว่าจะมีผลต่อการเกิดตัวแปรตอบสนอง มีขนาด $n \times (s+1)$, G คือ เมทริกซ์ของตัวแปรอธิบายที่คาดว่าจะมีผลต่อการเกิดค่าศูนย์ที่แท้จริง หรือเกิดค่าศูนย์จากการสุ่ม มีขนาด $n \times (t+1)$ โดยในงานวิจัยครั้งนี้ เมทริกซ์ X เท่ากับเมทริกซ์ G เนื่องจากข้อมูลเป็นข้อมูลชุดเดียวกัน, β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่มีขนาด $(s+1)\times 1$ และ γ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่มีขนาด $(t+1)\times 1$ ซึ่งค่าเฉลี่ย และค่าความแปรปรวนของ Y_i คือ

$$E(Y_i) = (1-p_i)\lambda_i$$
$$Var(Y_i) = (1-p_i)\lambda_i(1+p_i\lambda_i)$$

จากเงื่อนไข $\log(\lambda_i) = X_i\beta$ ซึ่งในแต่ละ $i = 1,2,3,...,n$

จัดได้อีกรูป คือ $\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_s x_{is}$; $i =$ 1,2,3,..., n

หรือ

$$\begin{bmatrix} \log\lambda_1 \\ \log\lambda_2 \\ \vdots \\ \log\lambda_n \end{bmatrix}_{(n\times 1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1s} \\ 1 & x_{21} & x_{22} & \cdots & x_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ns} \end{bmatrix}_{n\times(s+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_s \end{bmatrix}_{(s+1)\times 1}$$

จากเงื่อนไข $\text{logit}(p_i) = \log\left(\dfrac{p_i}{1-p_i}\right) = \underline{G}\gamma_i$

ตัวแบบการวิเคราะห์การถดถอยโลจิสติกที่มีตัวแปรอิสระมากกว่าหนึ่งตัวแปร สามารถเขียนสมการได้ดังนี้

$$\Pr ob(event) = \frac{e^z}{1+e^z}$$

โดย $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_t X_t$ และโอกาสของการไม่เกิดเหตุการณ์ จะหาได้จากสมการ

$$\Pr ob(noevent) = 1 - \Pr ob(event)$$

ตัวแบบการวิเคราะห์การถดถอยโลจิสติก สามารถเขียนในรูปของ odd ของการเกิดเหตุการณ์ได้ ซึ่ง odd ของการเกิดเหตุการณ์ หมายถึง อัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์กับโอกาสที่จะไม่เกิดเหตุการณ์ การเขียนตัวแบบโลจิสติกในรูปของ log ของ odd ซึ่งเรียกว่า logit สามารถเขียนได้ดังนี้

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \underline{G}\gamma_i$$

จัดได้อีกรูปคือ $\text{logit}(p_i) = \gamma_0 + \gamma_1 g_{i1} + \gamma_2 g_{i2} + ... + \gamma_i g_{it}$; $i =$ 1,2,3,..., n

หรือ

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_n) \end{bmatrix}_{(n\times 1)} = \begin{bmatrix} 1 & g_{11} & g_{12} & \cdots & g_{1t} \\ 1 & g_{21} & g_{22} & \cdots & g_{2t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_{n1} & g_{n2} & \cdots & g_{nt} \end{bmatrix}^T_{n\times(t+1)} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_t \end{bmatrix}_{(t+1)\times 1}$$

จากฟังก์ชันมวลความน่าจะเป็นของตัวแบบ ZIP สามารถหาลอกาลิทึมของฟังก์ชันล็อกภาวะน่าจะเป็น (Log-Likelihood Function) ได้ดังนี้

ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของตัวแบบ ZIP คือ

$$L(\gamma, \beta; Y) = \prod_{y_i=0} \Pr(Y = y_i) + \prod_{y_i>0} \Pr(Y = y_i)$$

เขียนเป็นฟังก์ชันลอกาลิทึมภาวะน่าจะเป็น ได้ดังนี้

$$\log L(\gamma, \beta; Y) = \sum_{y_i=0} \log \Pr(Y = y_i) + \sum_{y_i>0} \log \Pr(Y = y_i)$$

$$= \sum_{y_i=0} \log\left\{p_i + (1-p_i)e^{-\lambda_i}\right\} + \sum_{y_i>0} \log\left\{(1-p_i)\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right)\right\} \quad .....3$$

แต่เนื่องจาก $\log(\lambda_i) = \underline{X}_i\beta$ ซึ่งก็คือ $\lambda_i = e^{X_i\beta}$ ดังนั้นสำหรับแต่ละ $i = 1,2,3,...,n$ เมื่อ X_i คือสมาชิกในแถวที่ i ของ X และในทำนองเดียวกัน สำหรับ $\log\left(\frac{p}{1-p}\right) = G\gamma$ แล้ว $\frac{p}{1-p} = e^{G\gamma}$

และพบว่า $p_i = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}$ สำหรับ $i = 1,2,3,...,n$ เมื่อ G_i คือสมาชิกในแถวที่ i ของ G

จากสมการที่ 3

$$p_i + (1-p_i)e^{-\lambda_i} = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \left(1 - \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}\right)e^{-e^{X_i\beta}}$$

$$= \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \left(\frac{1+e^{G_i\gamma} - e^{G_i\gamma}}{1+e^{G_i\gamma}}\right)e^{-e^{X_i\beta}}$$

$$= \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \frac{e^{-e^{X_i\beta}}}{1+e^{G_i\gamma}}$$

$$= \frac{1}{\left(1+e^{G_i\gamma}\right)}\left\{e^{G_i\gamma} + e^{-e^{X_i\beta}}\right\}$$

พิจารณาเมื่อ $\log \Pr(Y_i = 0) = \log\left\{p_i + (1-p_i)e^{-\lambda_i}\right\}$

$$= \log\left\{\frac{1}{\left(1+e^{G_i\gamma}\right)}\left\{e^{G_i\gamma} + e^{-e^{X_i\beta}}\right\}\right\}$$

$$= \log\left(e^{G_i\gamma} + e^{-e^{X_i\beta}}\right) - \log\left(1+e^{G_i\gamma}\right)$$

พิจารณาเมื่อ $\log \Pr(Y_i > 0) = \log\left\{(1-p_i)\left(\frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y!}\right)\right\}$ เนื่องจาก $p_i = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}$ ดังนั้น

$$1 - p_i = 1 - \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} = \frac{1+e^{G_i\gamma} - e^{G_i\gamma}}{1+e^{G_i\gamma}} = \frac{1}{1+e^{G_i\gamma}}$$

ทำให้ $\log \Pr(Y_i > 0) = \log\left\{(1-p_i)\left(\frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!}\right)\right\}$

$$= \log\left\{\left(\frac{1}{1+e^{G_i\gamma}}\right)\frac{e^{-e^{X_i\beta}}\left(e^{X_i\beta}\right)^{y_i}}{y_i!}\right\}$$

$$= \log\left(e^{-e^{X_i\beta}}\right) + y_i\log\left(e^{X_i\beta}\right) - \log\left(1+e^{G_i\gamma}\right) - \log\left(y_i!\right)$$

$$= -e^{X_i\beta} + y_i X_i\beta - \log\left(1+e^{G_i\gamma}\right) - \log\left(y_i!\right)$$

จากสมการที่ 3 สามารถเขียนเป็นฟังก์ชันลอกาลิทึมภาวะน่าจะเป็นของตัวแบบถดถอย ZIP ได้ดังนี้

$$\log L(\gamma,\beta;Y) = \sum_{y_i=0}\left[\log\left(e^{G_i\gamma} + e^{-e^{X_i\beta}}\right) - \log\left(1+e^{G_i\gamma}\right)\right] + \sum_{y_i>0}\left[-e^{X_i\beta} + y_i X_i\beta - \log\left(1+e^{G_i\gamma}\right) - \log\left(y_i!\right)\right]$$

$$= \sum_{y_i=0}\log\left(e^{G_i\gamma} + e^{-e^{X_i\beta}}\right) - \sum_{y_i=0}\log\left(1+e^{G_i\gamma}\right) + \sum_{y_i>0}\left(y_i X_i\beta - e^{X_i\beta}\right) - \sum_{y_i>0}\log\left(1+e^{G_i\gamma}\right) - \sum_{y_i>0}\log\left(y_i!\right)$$

$$= \sum_{y_i=0}\log\left(e^{G_i\gamma} + e^{-e^{X_i\beta}}\right) + \sum_{y_i>0}\left(y_i X_i\beta - e^{X_i\beta}\right) - \left\{\sum_{y_i=0}\log\left(1+e^{G_i\gamma}\right) + \sum_{y_i>0}\log\left(1+e^{G_i\gamma}\right)\right\} - \sum_{y_i>0}\log\left(y_i!\right)$$

$$= \sum_{y_i=0}\log\left(e^{G_i\gamma} + e^{-e^{X_i\beta}}\right) + \sum_{y_i>0}\left(y_i X_i\beta - e^{X_i\beta}\right) - \sum_{i=1}^{n}\log\left(1+e^{G_i\gamma}\right) - \sum_{y_i>0}\log\left(y_i!\right)$$

ดังนั้น ฟังก์ชันลอกาลิทึมภาวะน่าจะเป็น ของตัวแบบการถดถอย ZIP เขียนได้ดังนี้

$$\log L(\gamma,\beta;Y)=\sum_{y_i=0}\log\left(e^{G_i\gamma}+e^{-e^{X_i\beta}}\right)+\sum_{y_i>0}\left(y_iX_i\beta-e^{X_i\beta}\right)-\sum_{i=1}^{n}\log\left(1+e^{G_i\gamma}\right)-\sum_{y_i>0}\log(y_i!) \quad \text{..... 4}$$

เรียกสมการนี้ว่า Incomplete Log-Likelihood Equation ในการประมาณค่าสัมประสิทธิ์ถดถอย γ และ β เพื่อให้ได้ตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator) โดยการหาอนุพันธ์สมการที่ 4 เทียบกับทั้ง γ และ β แล้วกำหนดให้ผลของการหาอนุพันธ์เทียบเท่ากับศูนย์ ซึ่งสมการที่ได้ไม่สามารถหาตัวประมาณผลลัพธ์ในรูป γ และ β ได้โดยง่าย การหาผลลัพธ์ซึ่งเป็นค่าประมาณของ γ และ β หาได้โดยใช้เทคนิคของ the Newton-Raphson Algorithm

เนื่องจากตัวประมาณ γ และ β ที่ได้จากวิธีข้างต้นเป็นตัวประมาณภาวะน่าจะเป็นสูงสุด (MLE) ซึ่งจะเป็นตัวประมาณที่มีการแจกแจงประมาณได้ด้วยการแจกแจงปกติที่มีค่าเฉลี่ย γ และ β และความแปรปรวนเท่ากับเมทริกซ์ข่าวสารสังเกตุผกผัน (the Inverse Observed Information Matrices) ปรากฏใน Lambert (1992) โดยในงานวิจัยชิ้นนี้ ผู้วิจัยจะคำนวณค่าประมาณของ γ และ β และทดสอบสมมติฐานต่างๆ เกี่ยวกับค่าของ γ และ β จากโปรแกรมสำเร็จ SAS

2) ตัวแบบการถดถอยทวินามนิเสธที่มีผลกระทบจากค่าศูนย์ (Zero-inflated Negative Binomial Regression Model : ตัวแบบการถดถอย ZINB) Greene [5]

กำหนดให้ ตัวแปรสุ่ม $Y=(y_1,y_2,...,y_n)'$ เป็นเวกเตอร์ของตัวแปรตอบสนองที่มีขนาดเท่ากับ n ที่แต่ละ $y_i;i=1,2,3,...,n$, เป็นอิสระต่อกัน เป็นตัวแปรสุ่มของตัวแปรตอบสนองที่มีการแจกแจงแบบ ZINB มีพารามิเตอร์ 3 ตัว ได้แก่ ตัวแรกคือ พารามิเตอร์ค่าเฉลี่ย $\left(\underline{\lambda}=(\lambda_1,\lambda_2,...,\lambda_n)'\right)$, $\lambda_i>0$ ซึ่งเป็นเวกเตอร์ของค่าคาดหวังของตัวแปรตอบสนองที่มีขนาด $n\times1$ ตัวที่สองคือ พารามิเตอร์ของความน่าจะเป็นที่จะเกิดศูนย์ $\left(\underline{p}=(p_1,p_2,...,p_n)'\right)$, $0\le p_i<1$, ซึ่งเป็น เวกเตอร์ของความน่าจะเป็นของสถานการณ์ที่จะเกิดศูนย์ที่มีขนาด $n\times1$ และตัวที่สามคือ พารามิเตอร์ดิสเพอชัน เป็นพารามิเตอร์ที่แสดงให้เห็นว่าค่าความแปรปรวนมีค่ามากกว่าหรือน้อยกว่าค่าเฉลี่ยมากน้อยเพียงใด $\left(\underline{\phi}=(\phi_1,\phi_2,...,\phi_n)'\right),\phi_i>0$ ซึ่งเวกเตอร์ของดิสเพอชัน (Dispersion Parameter) มีขนาด $n\times1$ โดยถ้าค่าพารามิเตอร์ดิสเพอชันมีค่ามากกว่า 1 แสดงว่าค่าความแปรปรวนมีค่ามากกว่าค่าเฉลี่ย ถ้าค่าพารามิเตอร์ดิสเพอชันมีค่าน้อยกว่า 1 แสดงว่าค่าความแปรปรวนมีค่าน้อยกว่าค่าเฉลี่ยซึ่งเป็นเหตุการณ์ที่พบได้ยาก เขียนการแจกแจงของตัวแปรสุ่มได้ด้วย $Y\sim \mathrm{ZINB}(\underline{\lambda},\underline{p},\underline{\phi})$ ซึ่ง

$y_i=0$ ด้วยความน่าจะเป็น $p_i, i=1,2,3,....n$

$y_i\neq0$ ด้วยความน่าจะเป็น $1-p_i, i=1,2,3,...,n$

ทำให้ฟังก์ชันมวลความน่าจะเป็นของ $Y \sim \text{ZINB}(\underline{\lambda}, \underline{p}, \underline{\phi})$ เป็น

$$\Pr(Y_i = y_i) = \begin{cases} p_i + (1-p_i)\left(\frac{\phi_i}{\phi_i + \lambda_i}\right)^{\phi_i} & ; y_i = 0 \\ (1-p_i)\frac{\Gamma(y_i+\phi_i)}{\Gamma(y_i+1)\Gamma(\phi_i)}\left(\frac{\phi_i}{\phi_i+\lambda_i}\right)^{\phi_i}\left(\frac{\lambda_i}{\phi_i+\lambda_i}\right)^{y_i} & ; y_i = 1,2,3,.. \end{cases} \quad .....5$$

โดยที่ $\lambda_i > 0$, $\phi_i > 0$, $0 \le p_i < 1$ และ $i = 1,2,3,...,n$

ที่มี $\underline{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)'$ และ $\underline{p} = (p_1, p_2, ..., p_n)'$ ที่สอดคล้องกับ

$$\log(\lambda) = \underline{X}\beta \quad \text{เมื่อต้องการประมาณค่าของตัวแปรตอบสนอง}$$

และ $$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \underline{G}\gamma \quad \text{เมื่อต้องการประมาณค่าความน่าจะเป็นที่จะเกิดศูนย์}$$

เมื่อ X คือ เมทริกซ์ของตัวแปรอธิบายที่คาดว่าจะมีผลต่อการเกิดตัวแปรตอบสนอง มีขนาด $n \times (s+1)$, G คือ เมทริกซ์ของตัวแปรอธิบายที่คาดว่าจะมีผลต่อการเกิดค่าศูนย์ที่แท้จริง หรือเกิดค่าศูนย์จากการสุ่ม มีขนาด $n \times (t+1)$ โดยในงานวิจัยครั้งนี้ เมทริกซ์ X เท่ากับเมทริกซ์ G เนื่องจากข้อมูลเป็นข้อมูลชุดเดียวกัน β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่มีขนาด $(s+1)\times 1$ และ γ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่มีขนาด $(t+1)\times 1$ ซึ่งค่าเฉลี่ย และค่าความแปรปรวนของ Y_i คือ

$$E(Y_i) = (1-p_i)\lambda_i$$
$$Var(Y_i) = (1-p_i)\lambda_i(1+\phi\lambda_i + p_i\lambda_i)$$

จากเงื่อนไข $\log(\lambda_i) = \underline{X}\beta_i$ ซึ่งในแต่ละ $i = 1,2,3,...,n$

จัดได้อีกรูปคือ $\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_s x_{is}$; $i = 1,2,3,...,n$

หรือ

$$\begin{bmatrix} \log\lambda_1 \\ \log\lambda_2 \\ \vdots \\ \log\lambda_n \end{bmatrix}_{(n\times 1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1s} \\ 1 & x_{21} & x_{22} & \cdots & x_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ns} \end{bmatrix}_{n\times(s+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_s \end{bmatrix}_{(s+1)\times 1}$$

จากเงื่อนไข $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \underline{G}\gamma_i$

โมเดลการวิเคราะห์การถดถอยโลจิสติกที่มีตัวแปรอิสระมากกว่าหนึ่งตัวแปร สามารถเขียนสมการได้ดังนี้

$$\Pr ob(event) = \frac{e^z}{1+e^z}$$

โดย $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_t X_t$

และโอกาสของการไม่เกิดเหตุการณ์ จะหาได้จากสมการ

$$\Pr ob(event) = 1 - \Pr ob(event)$$

โมเดลการวิเคราะห์การถดถอยโลจิสติก สามารถเขียนในรูปของ odd ของการเกิดเหตุการณ์ได้ ซึ่ง odd ของการเกิดเหตุการณ์ หมายถึง อัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์กับโอกาสที่จะไม่เกิดเหตุการณ์ การเขียนโมเดลโลจิสติกในรูปของ log ของ odd ซึ่งเรียกว่า logit สามารถเขียนได้ดังนี้

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \underline{G}\gamma_i$$

จัดได้อีกรูปคือ $\text{logit}(p_i) = \gamma_0 + \gamma_1 g_{i1} + \gamma_2 g_{i2} + ... + \gamma_t g_{it}$; $i = 1,2,3,...,n$

หรือ

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_n) \end{bmatrix}_{(n\times 1)} = \begin{bmatrix} 1 & g_{11} & g_{12} & \cdots & g_{1t} \\ 1 & g_{21} & g_{22} & \cdots & g_{2t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_{n1} & g_{n2} & \cdots & g_{nt} \end{bmatrix}^T_{n\times(t+1)} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_t \end{bmatrix}_{(t+1)\times 1}$$

จากฟังก์ชันมวลความน่าจะเป็นของตัวแบบ ZINB สามารถหาลอกาลิทึมของฟังก์ชันล็อกภาวะน่าจะเป็น (Log-Likelihood Function) ได้ดังนี้

ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของตัวแบบ ZINB คือ

$$L(\gamma, \beta; \underline{Y}) = \prod_{y_i=0} \Pr(Y = y_i) + \prod_{y_i>0} \Pr(Y = y_i)$$

เขียนเป็นฟังก์ชันฟังก์ชันล็อกความน่าจะเป็น (Log-Likelihood Function) ได้ดังนี้

$$\log L(\gamma, \beta; \underline{Y}) = \sum_{y_i=0} \log \Pr(Y = y_i) + \sum_{y_i>0} \log \Pr(Y = y_i)$$

$$= \sum_{y_i=0} \log\left\{ p_i + (1-p_i)\left(\frac{\phi}{\phi+\lambda_i}\right)^{\phi} \right\}$$

$$+ \sum_{y_i>0} \log\left\{ (1-p_i)\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\phi+\lambda_i}\right)^{\phi}\left(\frac{\lambda_i}{\phi+\lambda_i}\right)^{y_i} \right\} \quad \text{..... 6}$$

แต่เนื่องจาก $\log(\lambda_i) = \underline{X}_i\beta$ ซึ่งก็คือ $\lambda_i = e^{X_i\beta}$ ดังนั้นสำหรับแต่ละ $i = 1,2,3,...,n$

เมื่อ X_i คือสมาชิกในแถวที่ i ของ X และในทำนองเดียวกัน สำหรับ $\log\left(\frac{p}{1-p}\right) = G\gamma$ แล้ว $\frac{p}{1-p} = e^{G\gamma}$

และพบว่า $p_i = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}$ สำหรับ $1,2,3,...,n$ เมื่อ G_i คือสมาชิกในแถวที่ i ของ G

จากสมการที่ 6

$$p_i + (1-p_i)\left(\frac{\phi_i}{\phi_i+\lambda_i}\right)^{\phi_i} = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \left(1-\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}\right)\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}$$

$$= \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \left(\frac{1+e^{G_i\gamma}-e^{G_i\gamma}}{1+e^{G_i\gamma}}\right)\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}$$

$$= \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \frac{1}{1+e^{G_i\gamma}}\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}$$

พิจารณาเมื่อ $\log \Pr(Y_i = 0) = \log\left\{ p_i + (1-p_i)\left(\frac{\phi_i}{\phi_i+\lambda_i}\right)^{\phi_i} \right\}$

$$= \log\left\{ \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \left(1-\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}\right)\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i} \right\}$$

$$= \log\left\{ \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \frac{1}{1+e^{G_i\gamma}}\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i} \right\}$$

พิจารณาเมื่อ $\log \Pr(Y_i > 0) = \log\left\{ (1-p_i)\frac{\Gamma(y_i+\phi_i)}{\Gamma(y_i+1)\Gamma(\phi_i)}\left(\frac{\phi_i}{\phi_i+\lambda_i}\right)^{\phi_i}\left(\frac{\lambda_i}{\phi_i+\lambda_i}\right)^{y_i} \right\}$

เนื่องจาก $p_i = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}$

ดังนั้น $$1-p_i = 1-\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} = \frac{1+e^{G_i\gamma}-e^{G_i\gamma}}{1+e^{G_i\gamma}} = \frac{1}{1+e^{G_i\gamma}}$$

ทำให้

$$\log\left\{(1-p_i)\frac{\Gamma(y_i+\phi_i)}{\Gamma(y_i+1)\Gamma(\phi_i)}\left(\frac{\phi_i}{\phi_i+\lambda_i}\right)^{\phi_i}\left(\frac{\lambda_i}{\phi_i+\lambda_i}\right)^{y_i}\right\}$$

$$= \log\left\{\left(\frac{1}{1+e^{G_i\gamma}}\right)\frac{\Gamma(y_i+\phi_i)}{\Gamma(y_i+1)\Gamma(\phi_i)}\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i}\right\}$$

$$= \log\left(\frac{1}{1+e^{G_i\gamma}}\right)+\log\left(\frac{\Gamma(y_i+\phi_i)}{\Gamma(y_i+1)\Gamma(\phi_i)}\right)+\log\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}+\log\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i}$$

$$= \log(1)-\log\left(1+e^{G_i\gamma}\right)+\log\left(\Gamma(y_i+\phi_i)\right)-\log\left(\Gamma(y_i+1)\Gamma(\phi_i)\right)+\log\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}+\log\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i}$$

$$= -\log\left(1+e^{G_i\gamma}\right)+\log\left(\Gamma(y_i+\phi_i)\right)-\log\left(\Gamma(y_i+1)\right)-\log\left(\Gamma(\phi_i)\right)+\log\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}+\log\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i}$$

จากสมการที่ 6 สามารถเขียนเป็นฟังก์ชันลอกาลิทึมล็อกภาวะน่าจะเป็นของตัวแบบถดถอย $ZINB$ ได้ดังนี้

$$\log L(\gamma,\beta;\underline{Y}) = \sum_{y_i=0}\left[\log\left\{\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}+\frac{1}{1+e^{G_i\gamma}}\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}\right\}\right]$$

$$+\sum_{y_i>0}\left\{-\log\left(1+e^{G_i\gamma}\right)+\log\left(\Gamma(y_i+\phi_i)\right)-\log\left(\Gamma(y_i+1)\right)-\log\left(\Gamma(\phi_i)\right)+\log\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}+\log\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i}\right\}$$

$$= \sum_{y_i=0}\left[\log\left\{\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}+\frac{1}{1+e^{G_i\gamma}}\left(\frac{\phi}{\phi+e^{X_i\beta}}\right)^{\phi}\right\}\right]-\sum_{y_i>0}\log\left(1+e^{G_i\gamma}\right)+\sum_{y_i>0}\log\left(\Gamma(y_i+\phi)\right)-\sum_{y_i>0}\log\left(\Gamma(y_i+1)\right)$$

$$-\sum_{y_i>0}\log\left(\Gamma(\phi)\right)+\sum_{y_i>0}\log\left(\frac{\phi}{\phi+e^{X_i\beta}}\right)^{\phi}+\sum_{y_i>0}\log\left(\frac{e^{X_i\beta}}{\phi+e^{X_i\beta}}\right)^{y_i}$$

ดังนั้น ฟังก์ชันลอกาลิทึมภาวะน่าจะเป็น ของตัวแบบการถดถอย $ZINB$ เขียนได้ดังนี้

$$\log L(\gamma,\beta;\underline{Y}) = \sum_{y_i=0}\left[\log\left\{\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}}+\frac{1}{1+e^{G_i\gamma}}\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}\right\}\right]-\sum_{y_i>0}\log\left(1+e^{G_i\gamma}\right)+\sum_{y_i>0}\log\left(\Gamma(y_i+\phi_i)\right)$$

$$-\sum_{y_i>0}\log\left(\Gamma(y_i+1)\right)-\sum_{y_i>0}\log\left(\Gamma(\phi_i)\right)+\sum_{y_i>0}\log\left(\frac{\phi_i}{\phi_i+e^{X_i\beta}}\right)^{\phi_i}+\sum_{y_i>0}\log\left(\frac{e^{X_i\beta}}{\phi_i+e^{X_i\beta}}\right)^{y_i} \quad \text{.....7}$$

เรียกสมการนี้ว่า Incomplete Log-Likelihood Equation ในการประมาณค่าสัมประสิทธิ์ถดถอย γ และ β เพื่อให้ได้ตัวประมาณภาวะน่าจะเป็นสูงสุด (MLE) ทำได้โดยการหาอนุพันธ์สมการ (7) เทียบกับทั้ง γ และ β เช่นเดียวกับตัวแบบ ZIP

3.5 ในการวิเคราะห์ปัจจัยที่มีผลต่อการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ ทั้งสองตัวแบบ คือ ตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB ด้วย Multivariate Analysis ที่ระดับนัยสำคัญ 0.1

3.6 คัดเลือกตัวแบบที่ดีที่สุดโดยใช้เกณฑ์ดังนี้

1) เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's Information Criterion : AIC)

ในปี ค.ศ. 1973 Akaike [6] ได้คิดค้นและนำเสนอเกณฑ์สารสนเทศนี้ โดยมีสมการดังนี้

$$AIC = -2 \log likelihood + 2k$$

โดยที่ k คือ ผลรวมของจำนวนพารามิเตอร์ ในการคัดเลือกตัวแบบที่ดีที่สุดจะพิจารณาจากตัวแบบที่มีค่า AIC น้อยที่สุด

2) เกณฑ์สารสนเทศของเบส์ (Bayesian Information Criterion : BIC)

ในปี ค.ศ. 1978 Gideon E. Schwarz [7] เป็นผู้คิดค้นและใช้วิธีการของเบส์ในการวิเคราะห์ จึงอาจเรียกอีกชื่อได้ว่า Schwarz Criterion หรือ Schwarz Information (SIC) โดยมีสมการดังนี้

$$BIC = -2 \log likelihood + k \ln(n)$$

โดยที่ n คือ ขนาดตัวอย่างและ k คือ ผลรวมของจำนวนพารามิเตอร์ ในการคัดเลือกตัวแบบที่ดีที่สุด จะพิจารณาจากตัวแบบที่มีค่า BIC น้อยที่สุด

โดยการวิเคราะห์ตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB กับการประยุกต์ใช้กับข้อมูลจำนวนครั้งในการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ จะประมวลผลข้อมูลด้วยโปรแกรมสำเร็จ Statistical Analysis System (SAS) เวอร์ชัน 9.4 คำสั่ง Proc Genmod

ผลการวิจัย

1. ผลการวิเคราะห์ลักษณะข้อมูลเบื้องต้นของข้อมูล

จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนมีตั้งแต่ 0 ถึง 6 ครั้ง โดยจำนวน 0 ครั้ง คิดเป็น 97.79% ของจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนทั้งหมด

เพศชายมีการทำประกันมากกว่าเพศหญิง คิดเป็น 63.62% และ 36.38% ตามลำดับ และเพศชายมีการเรียกร้องค่าสินไหมทดแทนมากกว่าเพศหญิง คิดเป็น 66.68% และ 33.12% ตามลำดับ

อายุของผู้เอาประกันที่มีการทำประกันมีตั้งแต่อายุ 18 ถึง 82 ปี โดยอายุที่ทำประกันมากที่สุดคืออายุ 46 ปี คิดเป็น 15.22% ของจำนวนผู้ที่ทำประกันทั้งหมด ในขณะที่อายุของผู้เอาประกันที่มีการเรียกร้องค่าสินไหมทดแทนมากที่สุดคืออายุ 44 ปี คิดเป็น 17.72% ของผู้เอาประกันที่เรียกร้องค่าสินไหมทดแทน

อายุของรถยนต์ที่มีการทำประกันมีตั้งแต่รถยนต์อายุ 1 ปี ถึง 11 ปี โดยรถยนต์ที่มีการทำประกันมากที่สุดคือรถยนต์อายุ 2 ปี คิดเป็น 13.13% ของจำนวนรถยนต์ที่ทำประกันทั้งหมด ในขณะที่รถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทนมากที่สุดคือรถยนต์อายุ 11 ปี คิดเป็น 14.77% ของจำนวนรถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทน

ยี่ห้อรถยนต์ที่มีการทำประกันมีทั้งหมด 50 ยี่ห้อ โดยยี่ห้อที่มีการทำประกันมากที่สุดคือยี่ห้อ TOYOTA คิดเป็น 33.88% ของจำนวนรถยนต์ทั้งหมด ในขณะที่ยี่ห้อรถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทนมากที่สุดคือยี่ห้อ TOYOTA เช่นกัน คิดเป็น 40.51% ของจำนวนรถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทนทั้งหมด

รุ่นของรถยนต์ที่มีการทำประกันมีทั้งหมด 60 รุ่น โดยรุ่นที่มีการทำประกันมากที่สุดคือรถยนต์รุ่น TOYOTA COROLLA คิดเป็น 7.50% ของจำนวนรถยนต์ทั้งหมด ในขณะที่รุ่นของรถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทนมากที่สุดคือรุ่น TOYOTA SOLUNA คิดเป็น 10.76% ของจำนวนรถยนต์ที่มีการเรียกร้องค่าสินไหมทดแทนทั้งหมด กำหนดระดับนัยสำคัญสำหรับการสร้างตัวแบบจำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจเท่ากับ 0.1 โดยมีผลการประมวลข้อมูลดังนี้

2. ผลการวิเคราะห์ตัวแบบถดถอย

จากการประมวลผลข้อมูลโดยใช้โปรแกรม SAS เวอร์ชั่น 9.4 ด้วยคำสั่ง Proc Genmod โดยนำทุกปัจจัยไปวิเคราะห์พร้อมกันทั้งหมด (Multivariate Analysis) หากมีปัจจัยใดที่มีค่า p-value มากกว่า 0.1 ให้คัดออกจากสมการ จากนั้นจึงนำปัจจัยที่มีค่า p-value ที่น้อยกว่า 0.1 ไปวิเคราะห์ด้วยโปรแกรม SAS ใหม่อีกครั้ง เพื่อสร้างสมการถดถอย

2.1 ผลการวิเคราะห์ตัวแบบการถดถอย ZIP

ตารางที่ 1 ผลการประมาณค่าพารามิเตอร์จากตัวแบบการถดถอย ZIP

พารามิเตอร์				
	ค่าประมาณ	P-Value	ค่าประมาณ	P-Value
β_0	-0.3458	0.2855	-	-
β_1	-0.1130	0.2760	-	-
β_2	-0.0088	0.0443	-0.0124	0.0296
β_3	0.0824	< .0001	0.0915	0.0023
β_4	-0.0002	0.0268	-0.0087	0.0915
β_5	-0.0077	0.0678	-0.0093	0.0859
β_6	0.0163	0.0010	0.0546	0.0001

กำหนดให้ β_0 คือ ค่าคงที่ของสมการ และ β_1, β_2, β_3, β_4, β_5 และ β_6 คือสัมประสิทธิ์ของตัวแปรอธิบายเพศของผู้เอาประกัน (Gender) อายุของผู้เอาประกัน (Age) อายุของรถยนต์ (Vehicle Age) ขนาดตัวถังรถยนต์ (Vehicle Size) ยี่ห้อของรถยนต์ (Make) และรุ่นของรถยนต์ (Model) โดยเขียนเป็นสมการได้ดังนี้

$$\log(\lambda_i) = \beta_0 + \beta_1 Gender_i + \beta_2 Age_i + \beta_3 VehicleAge_i + \beta_4 VehicleSize_i + \beta_5 Make_i + \beta_6 Model_i$$

จากสมมติฐาน

$$H_0 : \beta_i = 0 \qquad ;i = 1,2,...,6$$
$$H_1 : \beta_i \neq 0 \qquad ;i = 1,2,...,6$$

โดยสมมติฐานจะปฏิเสธ H_0 เมื่อค่า p-value น้อยกว่า 0.1 ซึ่งหากพิจารณาจากค่า p-value แต่ละค่า สามารถเขียนสมการการถดถอย ZIP ได้ดังนี้

$$\log(\lambda_i) = -0.0124 Age_i + 0.0915 VehicleAge_i - 0.0087 VehicleSize_i - 0.0093 Make_i + 0.0546 Model_i$$

ตารางที่ 2 ผลการประมาณค่าพารามิเตอร์ที่มีผลกระทบจากค่าศูนย์จากตัวแบบการถดถอย ZIP

พารามิเตอร์				
	ค่าประมาณ	P-Value	ค่าประมาณ	P-Value
γ_0	2.0586	< .0001	2.9728	< .0001
γ_1	0.2853	0.0524	0.5386	< .0001
γ_2	0.0264	< .0001	0.1579	< .0001
γ_3	-0.0986	< .0001	-0.0087	< .0001
γ_4	0.0003	0.0415	0.0054	0.0533
γ_5	-0.0109	0.0165	-0.0096	0.0304
γ_6	0.0065	0.1948	-	-

กำหนดให้ γ_0 คือ ค่าคงที่ของสมการ และ γ_1, γ_2, γ_3, γ_4, γ_5 และ γ_6 คือสัมประสิทธิ์ของตัวแปรอธิบาย เพศของผู้เอาประกัน (Gender) อายุของผู้เอาประกัน (Age) อายุของรถยนต์ (Vehicle Age) ขนาดตัวถังรถยนต์ (Vehicle Size) ยี่ห้อของรถยนต์ (Make) และรุ่นของรถยนต์ (Model) ตามลำดับ โดยเขียนเป็นสมการได้ดังนี้

$$\log it(p_i) = \gamma_0 + \gamma_1 Gender_i + \gamma_2 Age_i + \gamma_3 VehicleAge_i + \gamma_4 VehicleSize_i + \gamma_5 Make_i + \gamma_6 Model_i$$

จากสมมติฐาน

$$H_0 : \gamma_i = 0 \qquad ; i = 1,2,...,6$$
$$H_1 : \gamma_i \neq 0 \qquad ; i = 1,2,...,6$$

โดยสมมติฐานจะปฏิเสธ H_0 เมื่อค่า p-value น้อยกว่า 0.1 ซึ่งหากพิจารณาจากค่า p-value แต่ละค่า สามารถเขียนสมการการถดถอย ZIP สำหรับข้อมูลที่มีผลต่อการเกิดค่าศูนย์ได้ดังนี้

$$\log it(p_i) = 2.9728 + 0.5386 Gender_i + 0.1579 Age - 0.0087 VehicleAge + 0.0054 VehicleSize_i - 0.0096 Make_i$$

2.2 ผลการวิเคราะห์ตัวแบบการถดถอย ZINB

ตารางที่ 3 ผลการประมาณค่าพารามิเตอร์จากตัวแบบการถดถอย ZINB

พารามิเตอร์				
	ค่าประมาณ	P-Value	ค่าประมาณ	P-Value
β_0	-0.9570	0.0845	-0.1824	0.0010
β_1	-0.1621	0.1553	-	-
β_2	-0.0128	0.0090	-0.1559	<.0001
β_3	0.0983	<.0001	0.2544	<.0001
β_4	-0.0002	0.0139	-0.0058	0.0048
β_5	-0.0085	0.1543	-	-
β_6	0.0190	0.0071	0.0795	<.0001

กำหนดให้ β_0 คือ ค่าคงที่ของสมการ และ β_1, β_2, β_3, β_4, β_5 และ β_6 คือสัมประสิทธิ์ของตัวแปรอธิบาย เพศของผู้เอาประกัน (Gender) อายุของผู้เอาประกัน (Age) อายุของรถยนต์ (Vehicle Age) ขนาดตัวถังรถยนต์ (Vehicle Size) ยี่ห้อของรถยนต์ (Make) และรุ่นของรถยนต์ (Model) ตามลำดับ โดยเขียนเป็นสมการได้ดังนี้

$$\log(\lambda_i) = \beta_0 + \beta_1 Gender_i + \beta_2 Age_i + \beta_3 VehicleAge_i + \beta_4 VehicleSize_i + \beta_5 Make_i + \beta_6 Model_i$$

จากสมมติฐาน

$$H_0 : \beta_i = 0 \qquad ;i = 1,2,...,6$$
$$H_1 : \beta_i \neq 0 \qquad ;i = 1,2,...,6$$

โดยสมมติฐานจะปฏิเสธ H_0 เมื่อค่า p-value น้อยกว่า 0.1 ซึ่งหากพิจารณาจากค่า p-value แต่ละค่า สามารถเขียนสมการการถดถอย ZINB ได้ดังนี้

$$\log(\lambda_i) = -0.1824 - 0.1559 Age_i + 0.2544 VehicleAge_i - 0.0058 VehicleSize_i + 0.0795 Model_i$$

ตารางที่ 4 ผลการประมาณค่าพารามิเตอร์ที่มีผลกระทบจากค่าศูนย์จากตัวแบบการถดถอย ZINB

พารามิเตอร์				
	ค่าประมาณ	P-Value	ค่าประมาณ	P-Value
γ_0	1.5194	0.0100	0.8742	0.0010
γ_1	0.3140	0.0734	0.6790	0.0022
γ_2	0.0292	< .0001	0.0119	< .0001
γ_3	-0.1012	0.0003	-0.1417	< .0001
γ_4	0.0003	0.0822	0.0001	0.0434
γ_5	-0.0111	0.0336	-0.2843	0.0915
γ_6	0.0071	0.2284	-	-

กำหนดให้ γ_0 คือ ค่าคงที่ของสมการ และ γ_1, γ_2, γ_3, γ_4, γ_5 และ γ_6 คือสัมประสิทธิ์ของตัวแปรอธิบาย เพศของผู้เอาประกัน (Gender) อายุของผู้เอาประกัน (Age) อายุของรถยนต์ (Vehicle Age) ขนาดตัวถังรถยนต์ (Vehicle Size) ยี่ห้อของรถยนต์ (Make) และรุ่นของรถยนต์ (Model) ตามลำดับ โดยเขียนเป็นสมการได้ดังนี้

$$\log it(p_i) = \gamma_0 + \gamma_1 Gender_i + \gamma_2 Age_i + \gamma_3 VehicleAge_i + \gamma_4 VehicleSize_i + \gamma_5 Make_i + \gamma_6 Model_i$$

จากสมมติฐาน

$$H_0 : \gamma_i = 0 \qquad ; i = 1,2,...,6$$
$$H_1 : \gamma_i \neq 0 \qquad ; i = 1,2,...,6$$

โดยสมมติฐานจะปฏิเสธ H_0 เมื่อค่า p-value น้อยกว่า 0.1 ซึ่งหากพิจารณาจากค่า p-value แต่ละค่าสามารถเขียนสมการการถดถอย ZINB สำหรับข้อมูลที่มีผลต่อการเกิดค่าศูนย์ ได้ดังนี้

$$\log it(p_i) = 0.8742 + 0.6790 Gender_i + +0.0119 Age - 0.1417 VehicleAge + 0.0001 VehicleSize_i - 0.2843 Make_i$$

ตารางที่ 5 การเปรียบเทียบความเหมาะสมของตัวแบบการถดถอย ZIP และตัวแบบการถดถอย ZINB

เกณฑ์ในการพิจารณา	ตัวแบบการถดถอย	
	ZIP	ZINB
ค่า Log Likelihood	-2,685.8329	-2,479.2851
ค่า AIC	5,383.6658	4,970.5702
ค่า BIC	5,431.5053	5,018.4097

ในการคัดเลือกตัวแบบที่ดีที่สุดจะพิจารณาจากค่า AIC และค่า BIC ที่ต่ำที่สุด และเนื่องจากจากตัวแบบทั้งสองค่า AIC และค่า BIC ที่พิจารณาได้จากตารางที่ 5 พบว่า ค่า AIC และค่า BIC ที่ได้จากตัวแบบการถดถอย ZINB ให้ค่าต่ำกว่าค่า AIC และค่า BIC ที่ได้จากตัวแบบการถดถอยที่ได้จากตัวแบบการถดถอย ZIP ดังนั้นผู้วิจัยจึงเลือกสมการการถดถอยที่ได้จากตัวแบบ ZINB เพื่อใช้กับข้อมูลชุดนี้

สรุปและอภิปรายผล

การวิเคราะห์ข้อมูลกับตัวแบบ ZIP และตัวแบบ ZINB ด้วยเกณฑ์สารสนเทศ AIC และ BIC พบว่า ตัวแบบที่ถูกเลือกคือตัวแบบ ZINB โดยมีปัจจัยที่มีผลต่อจำนวนครั้งในการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจสำหรับงานวิจัยชิ้นนี้ คือ อายุของผู้เอาประกัน อายุของรถยนต์ ขนาดตัวถังรถยนต์ และรุ่นของรถยนต์

ในส่วนของปัจจัยที่มีผลต่อการเกิดค่าศูนย์ พบว่า ตัวแบบ ZINB มีปัจจัยเพศของผู้เอาประกัน อายุของผู้เอาประกัน ขนาดตัวถังรถยนต์ และยี่ห้อของรถยนต์ที่มีผลต่อการเกิดค่าศูนย์

ข้อเสนอแนะ

1. งานวิจัยนี้ใช้ข้อมูลการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจของบริษัทประกันภัยทั้งประเทศ ทำให้ไม่สามารถวิเคราะห์ข้อมูลสำคัญในบางส่วนได้ เช่น ทุนประกัน ส่วนลด เงินปันผล และเบี้ยลงโทษ เป็นต้น เนื่องจากในแต่ละบริษัทมีการตั้งเกณฑ์ของข้อมูลเหล่านี้แตกต่างกัน

2. ข้อมูลที่ใช้ในการวิจัยมาจากฐานข้อมูลของสมาคมวินาศภัยไทย จึงมีข้อจำกัดในการให้ข้อมูลเกี่ยวกับตัวแปรอธิบาย ซึ่งอาจส่งผลต่อความถูกต้องและแม่นยำของตัวแบบ

3. จากการแบ่งกลุ่มของยี่ห้อรถยนต์ พบว่าสามารถแบ่งได้ถึง 50 กลุ่ม ซึ่งอาจเป็นจำนวนที่มากเกินไปหากนำมาใช้จริง จึงควรรวมกลุ่มของรถยนต์เหล่านี้เข้าด้วยกันอีก โดยประสานงานกับฝ่ายรับประกันภัยและฝ่ายสินไหมทดแทนถึงความสมเหตุสมผลต่อไป

เอกสารอ้างอิง

[1] Karen C.H. Yip.; and Kelvin K.W Yau. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance : Mathematics and Economics.* 33(2): 153-163.

[2] Noriszura Ismail.; and Hossein Zamani. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. in *Casualty Actuarial Society E-Forum*. n.p.

[3] Andrew Leung. (2011). Report on motor classification for voluntary motor insurance. *The Insurance Premium Bureau.* 24: 6-16.

[4] Lambert, D. (1992). Zero-inflated Poisson Regression with an application to defects in manufacturing. *Technometrics.* 34: 1-14.

[5] Greene WH. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. in *Working paper*. Department of Economics, Stern School of business. New York: New York University.

[6] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov.; and F. Csaki). Akademiai Kiado, Budapest. 267-281.

[7] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statisti.* 6: 461-464.