บทความวิจัย

ตัวแบบเชิงเส้นวางนัยทั่วไปสำหรับการศึกษาติดตามระยะยาวของจำนวนการเรียกค่าสินไหมทดแทนการประกันภัยรถยนต์ในกรุงเทพมหานคร

ศิริญญา ธีระอนันต์ชัย และ ลี่ลี อิงศรีสว่าง*

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแบบที่เหมาะสมสำหรับจำนวนการเรียกค่าสินไหมทดแทนการประกันภัยรถยนต์ในกรุงเทพมหานครที่มีการเก็บข้อมูลซ้ำตั้งแต่ปี พ.ศ. 2544-2548 โดยใช้ข้อมูลของผู้ทำประกันภัยรถยนต์ที่มีจำนวนการเรียกค่าสินไหมทดแทนอย่างน้อยหนึ่งครั้งในระยะเวลาติดตาม 5 ปี ทั้งหมด 3,635 กรมธรรม์ ที่ได้จากกรมการประกันภัย ด้วยตัวแบบ Generalized Estimating Equations (GEE) เมื่อกำหนดโครงสร้างความสัมพันธ์ของข้อมูลเป็นแบบ First-order Autoregressive (AR(1)) และ Compound Symmetry (CS) ตามลำดับ และตัวแบบผสมเชิงเส้นวางนัยทั่วไป (Generalized Linear Mixed Models, GLMMs) เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของข้อมูลเป็นแบบ AR(1) และ CS ตามลำดับ พร้อมทั้งพิจารณาตัวแบบที่เหมาะสม โดยตัวแปรตาม คือ จำนวนการเรียกค่าสินไหมทดแทนซึ่งมีการแจกแจงแบบปัวส์ซง และตัวแปรอิสระประกอบด้วยค่าสินไหมทดแทน อายุรถยนต์ เพศ กลุ่มอายุผู้ทำประกันภัย กลุ่มรถยนต์และขนาดเครื่องยนต์ ผลการศึกษาพบว่า ตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์เป็นแบบ AR(1) มีความเหมาะสมสำหรับข้อมูลมากกว่ารูปแบบ CS ด้วยค่าสถิติ Pearson Chi-square of residual/DF ที่มีค่าเข้าใกล้ 1 มากกว่าเล็กน้อยคือ 0.64 และ 0.63 ตามลำดับ ส่วนตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมเป็นแบบ AR(1) มีความเหมาะสมสำหรับข้อมูลมากกว่ารูปแบบ CS ด้วยค่าสถิติ Generalized Chi-Square/DF เท่ากับ 0.18 และ 0.15 ตามลำดับ เมื่อเปรียบเทียบตัวแบบ GEE และตัวแบบ GLMMs ในภาพรวมของค่าเฉลี่ยประชากรจะให้ค่าร้อยละของการทำนายถูกต้อง (R_p^2) ใกล้เคียงกัน ประมาณร้อยละ 53 แต่เมื่อพิจารณาตัวแบบ GLMMs ที่มีการเพิ่มเทอมอิทธิพลค่าคงที่สุ่มที่แสดงถึงความแตกต่างของแต่ละหน่วยศึกษา พบว่า ตัวแบบ GLMMs ที่มีโครงสร้างความแปรปรวนร่วมเป็นแบบ AR(1) ให้ประสิทธิภาพในการทำนายจำนวนการเรียกค่าสินไหมทดแทนมีความถูกต้องสูงขึ้นเป็นร้อยละ 76.28 และปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 ได้แก่ ค่าสินไหมทดแทน อายุรถยนต์ กลุ่มอายุผู้ทำประกันภัย อาชีพ และเพศ

คำสำคัญ: ตัวแบบผสมเชิงเส้นวางนัยทั่วไป ตัวแบบ generalized estimating equations (GEE), first-order autoregressive (AR(1)), compound symmetry (CS), ค่าคงที่สุ่ม จำนวนการเรียกค่าสินไหมทดแทน

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

*ผู้นิพนธ์ประสานงาน, e-mail: fscilli@ku.ac.th

Generalized Linear Models for Longitudinal Study of Car Insurance Claims in Bangkok

Sirinya Teeraananchai and Lily Ingsrisawang*

ABSTRACT

The goal of this study was to model a number of car insurance claims for car policyholders in Bangkok during the five-year insurance period of 2001-2005. The data used in this study came from the Department of Insurance in Thailand which consisted of 3,635 observations with at least one claim count in the 5-year period. The methodologies of GEE and GLMMs modeling approaches were applied by taking account of correlation and covariance structures of data such as First-order Autoregressive (AR(1)) and Compound Symmetry (CS), respectively. The appropriate model is suggested. The dependent variable was the claim counts with Poisson distribution while independent variables were indemnity, car-age, age, gender, occupation, car-group and engine size. The results showed that the GEE model with AR(1) correlation structure was more appropriate than the model with CS structure, as indicated by the values of Pearson Chi-square of residual/DF 0.64 and 0.63, respectively. On the other hand, the GLMMs model with AR(1) covariance structure was also more appropriate than the model with CS structure, as indicated by the values of Generalized Chi-Square/DF 0.18 and 0.15, respectively. The GLMMs and GEE for population averaged models had the same performance for estimating the claim counts with the percentage of correct predictions (R_p^2) about 53%, but the GLMMs model for emphasizing on subject-specific with AR(1) structure and random intercept effect showed more efficiency in the percentage of correct prediction with 76.28%. The statistically significant factors at the 0.05 level consisted of indemnity, car-age, age-group, occupation and gender.

Keywords: generalized linear mixed models (GLMMs), generalized estimating equations (GEE), first-order autoregressive (AR(1)), compound symmetry (CS), random intercept, claim count

Department of Statistics, Faculty of Science, Kasetsart University

*Corresponding author, e-mail: fscilli@ku.ac.th

บทนำ

การประกันภัยรถยนต์มีความเกี่ยวข้องกับผู้ใช้รถยนต์ในการเดินทางเป็นอย่างมาก คือ มีส่วนช่วยในการคุ้มครองผู้ขับขี่ และช่วยลดความเสี่ยงในการเกิดอุบัติเหตุ ในทางสถิติได้มีการเก็บรวบรวมข้อมูลการประกันภัยรถยนต์ โดยข้อมูลที่น่าสนใจจากการเก็บข้อมูลคือ จำนวนการเรียกค่าสินไหมทดแทน ซึ่งเป็นข้อมูลที่บริษัทประกันไม่สามารถรู้ล่วงหน้าได้เกี่ยวกับการใช้รถยนต์ของเจ้าของกรมธรรม์ ดังนั้นการนำข้อมูลจำนวนการเรียกค่าสินไหมทดแทนมาสร้างตัวแบบทางสถิติสามารถนำมาใช้ประกอบการพิจารณาการกำหนดเบี้ยประกันในปีต่อไปได้ รวมทั้งสามารถพยากรณ์จำนวนการเรียกค่าสินไหมทดแทนล่วงหน้า จำนวนการเรียกค่าสินไหมทดแทนเป็นข้อมูลการนับ (count) มีการแจกแจงแบบปัวส์ซง (Poisson distribution) สามารถนำมาสร้างตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized Linear Models, GLMs) เพื่อใช้ในการทำนายข้อมูลจำนวนการเรียกค่าสินไหมทดแทนในอนาคต แต่ตัวแบบ GLMs เป็นตัวแบบสำหรับข้อมูลที่ค่าสังเกตเป็นอิสระกันและไม่มีการวัดซ้ำ [1, 2] แต่สำหรับข้อมูลการประกันภัยรถยนต์เป็นข้อมูลระยะยาวที่มีการวัดซ้ำในบุคคลเดียวกันซึ่งค่าสังเกตมีความสัมพันธ์กันจึงเหมาะสมกับตัวแบบ Generalized Estimating Equations (GEE) ที่พัฒนามาจากหลักการของตัวแบบ GLMs [3] สำหรับการสร้างตัวแบบภาพรวมของค่าเฉลี่ยประชากร (Population-Average Model, PA) โดยได้มีการนำโครงสร้างความสัมพันธ์ของข้อมูลที่เก็บซ้ำของหน่วยศึกษาหนึ่งๆ ที่เรียกเป็น Working Correlation Matrix เข้ามาพิจารณาในขั้นตอนของการประมาณค่าพารามิเตอร์ นอกจากนี้การพัฒนาตัวแบบข้อมูลที่มีการวัดซ้ำแบบติดตามระยะยาวสามารถใช้ตัวแบบผสมเชิงเส้นวางนัยทั่วไป (Generalized Linear Mixed Models, GLMMs) [4] มาเป็นตัวแบบที่ใช้อธิบายแต่ละหน่วยบุคคลหรือหน่วยศึกษา (Subject-Specific Models) โดยตัวแบบนี้พัฒนาจากตัวแบบภาพรวมค่าเฉลี่ยประชากรก่อน และยอมให้มีการเพิ่มเทอมอิทธิพลสุ่ม (random Effect) เข้ามาในตัวแบบเพื่อแสดงลักษณะเฉพาะของแต่ละบุคคล เช่น การยอมให้ค่าพื้นฐานหรือที่เรียกว่าค่าคงที่ (intercept) ของแต่ละบุคคลมีค่าแตกต่างกันได้แบบสุ่ม เรียกเป็น Random Intercept เพื่อให้อธิบายความแตกต่างของแต่ละบุคคลได้ชัดเจนกว่าการอธิบายผลในภาพรวมของค่าเฉลี่ยประชากร

ดังนั้น ในงานวิจัยนี้มีวัตถุประสงค์ คือ เพื่อหาตัวแบบทางสถิติที่เหมาะสมสำหรับข้อมูลจำนวนการเรียกค่าสินไหมทดแทนที่เป็นข้อมูลติดตามระยะยาวด้วยตัวแบบ Generalized Estimating Equations (GEE) เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ First-order Autoregressive (AR(1)) และ Compound Symmetry (CS) และตัวแบบผสมเชิงเส้นวางนัยทั่วไป (GLMMs) เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ First-order Autoregressive (AR(1)) และ Compound Symmetry (CS)

วิธีการศึกษา

การศึกษาวิจัยนี้ได้ใช้ข้อมูลทุติยภูมิที่ได้จากกรมการประกันภัย โดยเป็นข้อมูลการประกันภัยรถยนต์ภาคสมัครใจประเภท 1 แบบระบุชื่อผู้ขับขี่ รหัส 110 รถยนต์นั่งส่วนบุคคล ตั้งแต่ปี พ.ศ. 2544-2548 ที่เริ่มทำการประกันภัยตั้งแต่ปี 2544 ของผู้ทำประกันภัยคนเดียวกันที่มีการเรียกค่าสินไหมทดแทนอย่างน้อย 1 ครั้งในระยะเวลาติดตาม 5 ปี จำนวน 3,635 กรมธรรม์ เมื่อตัวแปรตาม คือ จำนวนการเรียกค่าสินไหมทดแทนมีการแจกแจงแบบปัวส์ซง ส่วนตัวแปรอิสระประกอบด้วยตัวแปรต่อไปนี้

- เพศ หมายถึง เพศของผู้ทำประกันภัยรถยนต์
- อาชีพผู้ทำประกันภัยรถยนต์ หมายถึง อาชีพของผู้ทำประกันภัยรถยนต์ [5] ซึ่งแบ่งเป็น 2 กลุ่ม คือ อาชีพที่มีความเสี่ยงมาก ได้แก่ นักข่าว ผู้รับเหมาก่อสร้าง วิศวกร และสถาปนิก เป็นต้น และอาชีพที่มีความเสี่ยงน้อย ได้แก่ ครู พยาบาล แพทย์ แม่บ้าน เจ้าของกิจการและอาจารย์ เป็นต้น
- อายุของผู้ทำประกันภัยรถยนต์ (ปี) หมายถึง อายุของผู้ทำประกันภัยรถยนต์ โดยมีการแบ่งช่วงอายุจากช่วงที่มีความเสี่ยงภัยน้อยไปยังช่วงที่มีความเสี่ยงภัยมากเป็น 4 ช่วงอายุ [6] คือ อายุเกิน 50 ปีขึ้นไป, อายุ 36-50 ปี, อายุ 25-35 ปี และอายุ 18-24 ปี
- ขนาดของเครื่องยนต์ (cc.) หมายถึง ขนาดของเครื่องยนต์ของรถที่ทำประกันภัยซึ่งแบ่งเป็น 2 ระดับ [7] คือ กลุ่มที่ 1 ขนาดเครื่องยนต์น้อยกว่าเท่ากับ 2000 cc. และกลุ่มที่ 2 ขนาดเครื่องยนต์มากกว่า 2000 cc.
- อายุรถยนต์ (ปี) หมายถึง อายุการใช้งานของรถยนต์ที่ทำการประกันภัย
- ค่าสินไหมทดแทนหรือค่าเสียหาย (บาท) หมายถึง ค่าเสียหายที่ได้รับในกรณีที่เกิดความเสียหายของตัวรถยนต์เมื่อทำการซ่อม
- กลุ่มรถยนต์ หมายถึง การจัดกลุ่มของรถยนต์ที่ทำประกันภัยไว้ แบ่งออกเป็น 5 กลุ่ม คือ กลุ่มที่ 1 กลุ่มที่ 2 กลุ่มที่ 3 กลุ่มที่ 4 และกลุ่มที่ 5 ตามกรมการประกันภัย

ทำการวิเคราะห์ด้วยตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS และตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS, โดยใช้โปรแกรม SAS ในการหาตัวแบบทางสถิติที่เหมาะสมสำหรับจำนวนการเรียกค่าสินไหมทดแทนมีขั้นตอนดังนี้

ขั้นตอน 1 ศึกษาลักษณะข้อมูลเบื้องต้นด้วยสถิติเชิงพรรณนาจากค่าร้อยละ ความถี่ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน และมัธยฐาน รวมทั้งตรวจสอบความสัมพันธ์ระหว่างตัวแปรที่ใช้ในการศึกษาด้วยการหาค่าสัมประสิทธิ์สหสัมพันธ์สเปียร์แมนระหว่างข้อมูลกลุ่มแบบอันดับด้วยกัน และวัดระดับความสัมพันธ์ระหว่างข้อมูลกลุ่มที่แทนลักษณะกับข้อมูลกลุ่มแบบอันดับด้วยค่าสถิติ Phi และ Cramer's V [8] พร้อมทั้งตรวจสอบการแจกแจงของข้อมูลเชิงปริมาณที่มีค่าต่อเนื่อง เช่น ค่าสินไหมทดแทนด้วย Histogram และแก้ปัญหาข้อมูลเชิงปริมาณของตัวแปรอิสระกรณีที่มีการแจกแจงแบบเบ้ขวา โดยการแปลงข้อมูลด้วยวิธี Box-Cox เพื่อทำให้ข้อมูลมีความแปรปรวนคงที่ (Stability Variances) จากการศึกษาข้อมูลเบื้องต้นพบว่าข้อมูลค่าสินไหมทดแทนมีการแจกแจงของข้อมูลเป็นลักษณะเบ้ขวา และมีความแปรปรวนไม่คงที่ เมื่อนำมาทำการประมาณค่าพารามิเตอร์ในตัวแบบจะทำให้เกิดปัญหาความคลาดเคลื่อนของค่าพารามิเตอร์มีค่าต่ำเกินไป จึงได้ทำการแปลงค่าตัวแปรค่าสินไหมทดแทนที่มีการแจกแจงแบบเบ้ขวาด้วยวิธี Box-Cox [9] โดยอาศัยหลักการแปลงข้อมูลดังนี้

$$X_j^{(\lambda)} = \begin{cases} \log_e(X_j) & ; \lambda = 0 \\ \dfrac{(X_j^{\lambda} - 1)}{\lambda} & ; \lambda \neq 0 \end{cases}$$

เมื่อ X_j คือ ตัวแปรอิสระตัวที่ j ; j = 1, 2,..., k และ λ คือ ค่าพารามิเตอร์การแปลง ตัวแปรค่าสินไหมทดแทนมีบางค่าเป็นศูนย์ ทำให้ไม่สามารถทำการแปลงข้อมูลได้หากกำหนด $\lambda = 0$ จึงได้ทดลองเปลี่ยนค่า λ ที่เหมาะสม ซึ่งในการศึกษานี้ได้ทำการทดลองจนได้ค่าการแปลงข้อมูลที่เหมาะสม คือ $\lambda = 0.1$ เรียกตัวแปรนี้เป็นค่าสินไหมทดแทนที่แปลงค่า ซึ่งจะใช้เป็นตัวแปรอิสระในตัวแบบต่อไป และทำการแบ่งข้อมูลเป็น 2 ชุด โดยใช้หลักวิธี Cross-Validation [10] นำมาพัฒนาตัวแบบทางสถิติสำหรับจำนวนการเรียกค่าสินไหมทดแทน ด้วยวิธีการสุ่มอย่างง่าย คือ ทำการสุ่มขนาดตัวอย่างร้อยละ 70 ของข้อมูลทั้งหมดเป็นชุดข้อมูลฝึกสอนสำหรับนำมาพัฒนาตัวแบบทางสถิติแทนการใช้ข้อมูลทั้งหมด และที่เหลือร้อยละ 30 เป็นชุดข้อมูลทดสอบเป็นข้อมูลสำหรับนำมาทำนายจำนวนการเรียกค่าสินไหมทดแทน โดยใช้ตัวแบบทางสถิติที่ประมาณค่าพารามิเตอร์ได้จากชุดข้อมูลฝึกสอน โดยข้อมูลที่ทำการแบ่งแล้วต้องนำมาตรวจสอบลักษณะและการแจกแจงของข้อมูลทั้งสองชุดจะต้องมีลักษณะเช่นเดียวกับข้อมูลทั้งหมด

ขั้นตอน 2 นำข้อมูลฝึกสอนมาพัฒนาตัวแบบทางสถิติ เริ่มจากการวิเคราะห์ตัวแปรเดียว (Univariate Analysis) ระหว่างตัวแปรตามจำนวนการเรียกค่าสินไหมทดแทนกับตัวแปรอิสระทีละตัวกับตัวแบบต่อไปนี้

- ตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS พิจารณาจากตัวสถิติทดสอบ Wald
- ตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS พิจารณาจากตัวสถิติทดสอบ t

ต่อจากนั้นตัวแปรอิสระที่ให้ p-value น้อยกว่า 0.20 ที่ได้จากแต่ละตัวแบบนำไปเป็นตัวแปรนำเข้าสำหรับการวิเคราะห์พหุตัวแปร (Multivariate Analysis) เพื่อคัดเลือกตัวแปรอิสระด้วยวิธี Forward Stepwise ในตัวแบบ GEE โดยพิจารณาจาก p-value ของตัวสถิติทดสอบ Generalized Score Statistic และในตัวแบบ GLMMs พิจารณาจาก p-value ของตัวสถิติทดสอบ G (Residual Log Pseudo-Likelihood Ratio Test Statistic) ตามเกณฑ์การคัดเลือกตัวแปรอิสระเข้าตัวแบบที่ระดับนัยสำคัญทางสถิติ 0.15 ($P_E = 0.15$) และเกณฑ์การคัดเลือกตัวแปรอิสระออกจากตัวแบบที่ระดับนัยสำคัญทางสถิติ 0.20 ($P_R = 0.20$) [11]

ขั้นตอน 3 นำตัวแปรอิสระที่ผ่านการคัดเลือกทั้งหมดจากชุดข้อมูลฝึกสอนมาสร้างตัวแบบทางสถิติและทำการทำนายจำนวนการเรียกค่าสินไหมทดแทน รวมทั้งนำตัวแบบที่ได้จากชุดข้อมูลฝึกสอนมาทำนายจำนวนการเรียกค่าสินไหมทดแทนในชุดข้อมูลทดสอบ และคำนวณค่าร้อยละของการทำนายถูกต้อง (R_p^2) ในชุดข้อมูลฝึกสอน และข้อมูลทดสอบว่ามีค่าร้อยละของการทำนายถูกต้องใกล้เคียงกันหรือไม่ ถ้ามีค่าใกล้เคียงกันหรือมากกว่านั้น แสดงว่าสามารถนำตัวแปรอิสระที่ผ่านการคัดเลือกจากข้อมูลฝึกสอนมาวิเคราะห์ในชุดข้อมูลทั้งหมดต่อไปได้

ขั้นตอน 4 หาตัวแบบสำหรับข้อมูลจำนวนการเรียกค่าสินไหมทดแทนที่เหมาะสมจากการวิเคราะห์ข้อมูลทั้งหมดโดยใช้ตัวแปรอิสระที่ผ่านการคัดเลือกจากชุดข้อมูลฝึกสอน จากตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS โดยพิจารณจากค่าสถิติ Chi-Square of Residuals/DF และค่า (R_p^2) และจากตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS โดยพิจารณาจากค่าสถิติ Generalized Chi-Square/DF ค่า Pseudo-AIC ค่า Pseudo-BIC และค่า (R_p^2)

การศึกษานี้ได้อาศัยทฤษฎีที่เกี่ยวข้องดังนี้

1. ตัวแบบ Generalized Estimating Equations (GEE)

ในการศึกษานี้ตัวแปรตามหรือ Y เป็นจำนวนการเรียกค่าสินไหมทดแทนในการประกันภัยที่มีการแจกแจงแบบปัวส์ซง วิธี GEE จะสมมติให้ y_{ij} คือ ค่าสังเกตของหน่วยศึกษาที่ i = 1,2,...,s จากการวัดครั้งที่ j = 1,2,...,t และมีความสัมพันธ์กับเวกเตอร์ค่าสังเกตของตัวแปรอิสระ $X_1,...,X_k$ ที่ได้จากการวัดครั้งที่ j คือ เมื่อ $\underline{x}'_{kj} = (1\ x_{ij},...,x_{pj})$ เมื่อ k = 1,2,...,p สามารถเขียนตัวแบบ GEE สำหรับการวิเคราะห์ Marginal Model ดังนี้ [12]

$$\log_e(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_p x_{pij} + e_{ij} \quad (1)$$

โดยประกอบด้วยลักษณะ 3 ข้อ คือ

1. ค่าเฉลี่ยของตัวแปรตาม (μ_{ij}) จะมีความสัมพันธ์กับตัวทำนายเชิงเส้นผ่านทางฟังก์ชันเชื่อมโยงของ Log Link ดังนี้

$$g(\mu_{ij}) = \log_e(\mu_{ij}) = \underline{x}'_{kj}\beta$$

2. ความแปรปรวนของตัวแปรตาม เมื่อถูกกำหนดด้วยค่าของตัวแปรอิสระจะขึ้นกับค่าเฉลี่ยของตัวแปรตามนั้น

จาก $Var(Y_{ij}) = \phi\, Var(\mu_{ij})$ เมื่อ $Var(\mu_{ij}) = \mu_{ij}$ และ $\phi = 1$ จะได้ $Var(Y_{ij}) = \mu_{ij}$

3. ความสัมพันธ์ระหว่างค่าสังเกตของหน่วยศึกษาเดียวกันที่ปรากฏใน Working Correlation Matrix จะสมมติให้อยู่ในรูปแบบของ AR(1) และ CS ดังนี้

3.1 AR(1) มีโครงสร้างดังนี้

$Corr(y_{ij},y_{ij+t}) = \rho^t$ fot t = 1,2, ..., n_i-j เขียนในรูปเมทริกซ์ คือ $\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$

ความสัมพันธ์ของค่าสังเกตภายในหน่วยศึกษาเดียวกันจะมีค่าลดลงตามระยะห่างของช่วงเวลาที่เก็บข้อมูลซ้ำ ซึ่งค่าความสัมพันธ์จะอยู่ในรูปของ ρ^t เมื่อ t คือ จำนวนช่วงเวลาก่อนหน้า (lag time) เมื่อ t = 1,2 ..., n_i-j

3.2 CS มีโครงสร้างดังนี้

$$\mathrm{Corr}(y_{ij},y_{ik}) = \begin{cases} 1 & ; j = k \\ \rho & ; j \neq k \end{cases} \quad \text{เขียนในรูปเมทริกซ์ คือ} \begin{pmatrix} 1 & \rho & \rho \dots \rho \\ \rho & 1 & \rho \dots \rho \\ \rho & \rho & 1 \dots \rho \\ \vdots & \vdots & \vdots \ddots \vdots \\ \rho & \rho & \rho \dots 1 \end{pmatrix}$$

ความสัมพันธ์ของค่าสังเกตภายในหน่วยศึกษาเดียวกันจะมีค่าเท่ากัน เท่ากับ ρ ณ ช่วงเวลาที่ $j \neq k$ ซึ่งตัวแบบ GEE จะทำการประมาณค่าพารามิเตอร์ด้วยวิธี Quasi - Likelihood โดยใช้โปรแกรมสถิติ SAS ด้วยคำสั่ง PROC GENMOD [13]

2. ตัวแบบผสมเชิงเส้นวางนัยทั่วไป (Generalized Linear Mixed Models, GLMMs)

ตัวแบบผสมเชิงเส้นวางนัยทั่วไป (GLMMs) ประกอบด้วยเทอมอิทธิพลคงที่และอิทธิพลสุ่ม เมื่อตัวแปรตาม Y คือ จำนวนการเรียกค่าสินไหมทดแทน ที่มีการแจกแจงแบบปัวส์ซง สามารถเขียนตัวแบบ GLMMs ได้ดังนี้ [13]

$$\log_e(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{ij2} + \dots + \beta_p x_{ijk} + u_i + \varepsilon_{ij} \qquad (2)$$

โดยประกอบด้วยลักษณะ 3 ข้อ คือ

1. ค่าเฉลี่ยของตัวแปรตามแบบมีเงื่อนไขที่ขึ้นกับ u_i (μ_{ij}) จะมีความสัมพันธ์เชิงเส้นกับเทอมอิทธิพลคงที่และเทอมอิทธิพลสุ่ม ผ่านฟังก์ชันเชื่อมโยงในรูปแบบล็อกของตัวทำนายเชิงเส้น คือ

จาก $$E(\underline{Y}_{ij}| \underline{u}_i = X_{ij}\beta + Z_{ij}\underline{u}_i \qquad (3)$$

$$\eta_{ij} = X_{ij}\underline{\beta} + Z_{ij}\underline{u}_i \qquad \text{เมื่อ } \eta_{ij} = \log_e \{E(\underline{Y}_{ij}| \underline{u}_i)\}$$

จะได้ $$\log_e \{E(\underline{Y}_{ij}| \underline{u}_i)\} = X_{ij}\underline{\beta} + Z_{ij}\underline{u}_i \qquad (4)$$

โดยที่ Y_{ij} คือ ค่าสังเกตของตัวแปรตามสำหรับหน่วยศึกษาที่ i วัดครั้งที่ j มี X_{ij} คือ เมทริกซ์ของตัวแปรอิสระหรืออิทธิพลคงที่ที่มี k ตัวแปรสำหรับหน่วยศึกษาที่ i วัดครั้งที่ j โดยมีขนาดเมทริกซ์ n x p ; p = k +1, $\underline{\beta}$ คือเวกเตอร์ของพารามิเตอร์อิทธิพลคงที่ที่มีขนาด p x 1, Z_{ij} คือเมทริกซ์ของอิทธิพลสุ่มที่มี q ตัวแปรสำหรับหน่วยศึกษาที่ i วัดครั้งที่ j ที่มีขนาด n x q โดย $q \leq p$, $\underline{u}_i$ คือเวกเตอร์ของพารามิเตอร์อิทธิพลสุ่มของหน่วยศึกษาที่ i ที่มีขนาด q x 1 เมื่อ $\underline{u}_i \sim N(\underline{0},G)$ และ $\underline{\varepsilon}_i$ คือเวกเตอร์ของความคลาดเคลื่อนของหน่วยศึกษาที่ i เมื่อ $\underline{\varepsilon}_i \sim N(\underline{0},R_i)$

2. ตัวแปรตามมีการแจกแจงแบบปัวส์ซง สามารถหาค่า $Var(\underline{Y}_{ij}|\underline{u}_i)$ และค่า $E(\underline{Y}_{ij}|\underline{u}_i)$ คือ

$$Var(\underline{Y}_{ij}|\underline{u}_i) = E(\underline{Y}_{ij}|\underline{u}_i) \qquad \text{โดยที่ค่า } \phi = 1$$

เนื่องจากค่าเฉลี่ยและค่าความแปรปรวนของการแจกแจงแบบปัวส์ซงจะมีค่าเท่ากัน

3. เมื่อกำหนด u_i มีการแจกแจงปกติพหุตัวแปร (Multivariate Normal Distribution) คือ $\underline{u}_i \sim N(\underline{0},G)$ เมื่อ G เป็นเมทริกซ์ความแปรปรวนร่วม และเป็นอิสระกับ X_{ij} โดยเขียนโครงสร้างความแปรปรวนร่วมของตัวแปรตามได้ดังนี้

$$\begin{aligned} Var(\underline{Y}_i) &= Var(Z_i\underline{u}_i) + Var(\underline{\varepsilon}_i) \\ &= Z_iGZ_i' + R_i \end{aligned}$$

ในการศึกษานี้เป็นข้อมูลการวัดซ้ำจะสมมติให้โครงสร้างความแปรปรวนร่วมเป็นแบบ AR(1) เขียนในรูปเมทริกซ์ ดังนี้

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \cdots & \sigma^2 \end{pmatrix}$$

โครงสร้างความแปรปรวนร่วมแบบ AR(1) จะกำหนดให้แต่ละค่าสังเกตของหน่วยศึกษาที่ i มีความแปรปรวนเท่ากัน ส่วนความแปรปรวนร่วมระหว่างค่าสังเกตจะไม่เท่ากันสำหรับบุคคลเดียวกัน แต่จะลดลงเข้าใกล้ศูนย์เมื่อช่วงห่างเวลาของการวัดซ้ำข้อมูลมากขึ้น

และมีโครงสร้างความแปรปรวนร่วมเป็นแบบ CS เขียนในรูปเมทริกซ์ ดังนี้

$$Ri = Var(\varepsilon_i) = \begin{pmatrix} \sigma^2+\sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2+\sigma_1 \end{pmatrix}$$

ค่าสังเกตจะมีค่าความแปรปรวนร่วมในแต่ละคู่เท่ากันสำหรับหน่วยศึกษาเดียวกัน

ในการวิเคราะห์ตัวแบบ GLMMs ได้ใช้คำสั่ง PROC GLIMMIX จากโปรแกรมสำเร็จรูป SAS [14] โดยทำการประมาณค่าพารามิเตอร์ในเทอมอิทธิพลคงที่ด้วยวิธี Restricted Pseudo-likelihood (REPL) และประมาณค่าในเทอมอิทธิพลสุ่มด้วยวิธี Maximum Likelihood [15]

ตารางที่ 1 การตรวจสอบความเหมาะสมของตัวแบบ

ตัวแบบ Generalized Estimating Equations (GEE)	**ตัวแบบผสมเชิงเส้นวางนัยทั่วไป (GLMMs)**
การตรวจสอบความเหมาะสมของตัวแบบ [16] 1. Pearson Chi-Square of Residuals/DF โดยค่าที่เหมาะสมจะมีค่าเข้าใกล้ 1 แต่ไม่ควรมากกว่า 1 เพราะอาจทำให้เกิดปัญหามีค่าความแปรปรวนมากเกินจริง (over-dispersion) เนื่องจากการแจกแจงแบบปัวส์ซงจะมีค่า ϕ=1 2. ค่าร้อยละของการทำนายถูกต้อง (Percent age of correct predictions) หรือ R_p^2 ใช้วัดความถูกต้องในการทำนายคำนวณได้ดังนี้ $R_p^2 = \frac{\text{จำนวนค่าสังเกตที่ทำนายถูก}}{\text{จำนวนค่าสังเกตทั้งหมด}} \times 100$	การตรวจสอบความเหมาะสมของตัวแบบ [16] 1. Generalized Chi-Square/DF โดยค่าที่เหมาะสมจะมีค่าเข้าใกล้ 1 แต่ไม่ควรมากกว่า 1 เพราะอาจทำให้เกิดปัญหามีค่าความแปรปรวนมากเกินจริง (over - dispersion) เนื่องจากการแจกแจงแบบปัวส์ซงจะมีค่า ϕ=1 2. ค่า Pseudo-AIC (Akaike's Information Criterion) และค่า Pseudo-BIC (Bayes' Information Criterion) สำหรับตรวจสอบความเหมาะสมของตัวแบบ GLMMs เมื่อเลือกใช้โครงสร้างความแปรปรวนร่วมต่างกัน ค่า Pseudo-AIC หรือ Pseudo-BIC ของตัวแบบที่มีค่าต่ำกว่าแสดงว่าตัวแบบนั้นเป็นตัวแบบที่เหมาะสมกับข้อมูลมากกว่า $\text{Pseudo-AIC} = -2l + 2d$ $\text{Pseudo-BIC} = -2l + d \log_e n$ เมื่อ l คือ Log Likelihood, d คือ จำนวนพารามิเตอร์ และ n คือ ขนาดตัวอย่าง 3. ค่าร้อยละของการทำนายถูกต้อง (Percentage of correct predictions) หรือ R_p^2 ใช้วัดความถูกต้องในการทำนายคำนวณได้ดังนี้ $R_p^2 = \frac{\text{จำนวนค่าสังเกตที่ทำนายถูก}}{\text{จำนวนค่าสังเกตทั้งหมด}} \times 100$

ผลการศึกษา

1. จากการศึกษาลักษณะข้อมูลเบื้องต้นเมื่อทำการแบ่งข้อมูลทั้งหมดเป็นข้อมูลฝึกสอน และข้อมูลทดสอบพบว่า ข้อมูลฝึกสอนและข้อมูลทดสอบมีลักษณะของข้อมูลเป็นแบบเดียวกับข้อมูลทั้งหมด จึงสามารถนำมาพัฒนาตัวแบบต่อไปได้

2. ผลการพัฒนาตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS และตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS จากการวิเคราะห์ตัวแปรเดียวและการวิเคราะห์หลายตัวแปรมีตัวแปรอิสระที่ผ่านการคัดเลือกเหมือนกันโดยพิจารณาจาก p-value ที่น้อยกว่า 0.20 คือ ค่าสินไหมทดแทนที่แปลงค่า อายุรถยนต์ กลุ่มอายุผู้ทำประกันภัย อาชีพ และเพศ ซึ่งได้ผลการวิเคราะห์ตัวแบบทางสถิติสำหรับข้อมูลจำนวนการเรียกค่าสินไหมทดแทนจากตัวแบบ GEE เมื่อมีโครงสร้างความสัมพันธ์เป็นแบบ AR(1) และ CS และตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS แสดงในตารางที่ 2 และ 3 ดังนี้

ตารางที่ 2 ผลวิเคราะห์ตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์เป็นแบบ AR(1) และ CS

ตัวแปรอิสระ	ตัวแบบ GEE			
	AR(1)		CS	
Fixed-effect Parameter	**Estimate (S.E.)**	**Z [p-value]**	**Estimate (S.E.)**	**Z [p-value]**
ค่าคงที่ (INTERCEPT)	-1.8074 (0.0530)	-34.08 [<.0001]*	-2.0178 (0.0612)	-32.97 [<.0001]*
ค่าสินไหมทดแทนที่แปลงค่า	0.1464 (0.0029)	50.23 [<.0001]*	0.1574 (0.0037)	42.97 [<.0001]*
อายุรถยนต์	0.0452 (0.0025)	18.15 [<.0001]*	0.0527 (0.0027)	19.20 [<.0001]*
อายุผู้ทำประกันภัย				
18-24 ปี	0.1113 (0.0835)	1.33 [0.1828]	0.1435 (0.0909)	1.58 [0.1143]
25-35 ปี	0.1656 (0.0350)	4.73 [<.0001]*	0.1737 (0.0397)	4.38 [<.0001]*
36-50 ปี	0.0787 (0.0329)	2.39 [0.0168]*	0.0824 (0.0376)	2.19 [0.0282]*
อาชีพ	0.3488 (0.0495)	7.04 [<.0001]*	0.3536 (0.0538)	6.58 [<.0001]*
เพศ	-0.0366 [0.0190]	-1.92 [0.0546]	-0.0428 (0.0209)	-2.05 [0.0404]*
Model fit Criteria	**AR(1)**		**CS**	
Pearson Chi-Square of residual	11682.680		11414.060	
Pearson Chi-Square of residual/DF	0.64		0.63	
R_p^2 (%) ของ $\hat{Y}$	52.32		52.05	

หมายเหตุ เมื่อ * มีระดับนัยสำคัญทางสถิติที่ระดับ 0.05
S.E. คือ ค่าความคลาดเคลื่อนมาตรฐานของค่าสัมประสิทธิ์การถดถอย
$\hat{Y}$ คือ จำนวนการเรียกค่าสินไหมทดแทนที่ทำนายจากตัวแบบ
R_p^2 คือ ค่าร้อยละของการทำนายถูกต้อง

จากตารางที่ 2 พบว่า การวิเคราะห์ด้วยตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) พบว่า ปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 คือ ค่าสินไหมทดแทนที่แปลงค่า อายุรถยนต์ กลุ่มอายุผู้ทำประกันภัย และอาชีพ ยกเว้นปัจจัยด้านเพศ ส่วนการวิเคราะห์ตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ CS ปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 คือ ค่าสินไหมทดแทน อายุของรถยนต์ กลุ่มอายุผู้ทำประกันภัย อาชีพ และเพศ เมื่อพิจารณาความเหมาะสมของตัวแบบพบว่า ค่าสถิติของการวิเคราะห์ตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) มีค่าสถิติ Pearson Chi-Square of residual/DF เท่ากับ 0.64 ส่วนรูปแบบ CS มีค่าสถิติ Pearson Chi-Square of residual/DF เท่ากับ 0.63 โดยรูปแบบ AR(1) มีค่า Pearson Chi-Square of residual และเข้าใกล้ค่าองศาความเป็นอิสระ (Degree of freedom) มากกว่ารูปแบบ CS และมีค่า Pearson Chi-Square of residual/DF ใกล้เคียงกัน แต่มีค่าน้อยกว่า 1 แสดงว่าไม่เกิดปัญหามีค่าความแปรปรวนเกินจริง และพิจารณาค่าร้อยละของความถูกต้องที่ได้จากการทำนายของ

ตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS มีค่าความถูกต้องเท่ากับร้อยละ 52.32 และ 52.05 ตามลำดับ ซึ่งมีค่าใกล้เคียงกัน

ตารางที่ 3 ผลวิเคราะห์ตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมเป็นแบบ AR(1) และ CS

ตัวแปรอิสระ	**ตัวแบบ GLMMs**			
	AR(1)		**CS**	
Fixed-effect Parameter	**Estimate (S.E.)**	**t [p-value]**	**Estimate (S.E.)**	**t [p-value]**
ค่าคงที่ (INTERCEPT)	-2.7130 (0.0467)	-58.07 [<.0001]*	-2.9525 (0.0458)	-64.54 [<.0001]*
ค่าสินไหมทดแทนที่แปลงค่า	0.1805 (0.0024)	75.19 [<.0001]*	0.1904 (0.0025)	76.69 [<.0001]*
อายุรถยนต์	0.0861 (0.0026)	32.82 [<.0001]*	0.0938 (0.0030)	31.55 [<.0001]*
อายุผู้ทำประกันภัย				
18-24 ปี	0.1390 (0.0862)	1.61 [0.1068]	0.1445 (0.0886)	1.63 [0.1028]
25-35 ปี	0.1703 (0.0310)	5.49 [<.0001]*	0.1755 (0.0323)	5.43 [<.0001]*
36-50 ปี	0.0744 (0.0288)	2.58 [0.0098]*	0.0751 (0.0301)	2.50 [0.0126]*
อาชีพ	0.3105 (0.0514)	6.04 [<.0001]*	0.3054 (0.0537)	5.69 [<.0001]*
เพศ	-0.0487 (0.0179)	-2.72 [0.0065]*	-0.0517 (0.0186)	-2.79 [0.0053]*
Random Subject Effect	**Estimate (S.E.)**		**Estimate (S.E.)**	
Subject 1	0.4135 (0.2593)		0.5648 (0.1876)	
Subject 2	-0.4155 (0.2105)		-0.4252 (0.1606)	
Model fit Criteria	**AR(1)**		**CS**	
Generalized Chi-Square	3345.96		2677.77	
Generalized Chi-Square/DF	0.18		0.15	
Pseudo-AIC	36630.43		40278.53	
Pseudo-BIC	36649.03		40290.93	
R_p^2 (%) of Y_u	76.28		79.36	
R_p^2 (%) ของ $\hat{Y}$	53.34		53.38	

หมายเหตุ เมื่อ * มีระดับนัยสำคัญทางสถิติที่ระดับ 0.05

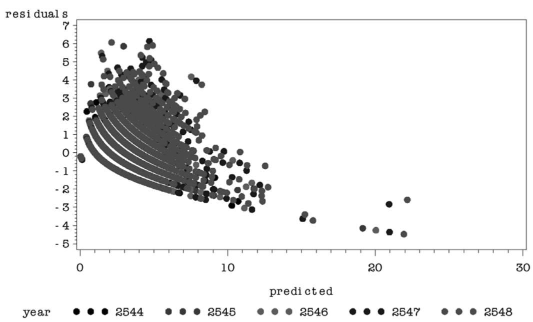
S.E. คือ ค่าความคลาดเคลื่อนมาตรฐานของค่าสัมประสิทธิ์การถดถอย

$\hat{Y}_u$ คือ จำนวนการเรียกค่าสินไหมทดแทนที่ทำนายจากตัวแบบ เมื่อมีเทอมอิทธิพลค่าคงที่สุ่ม

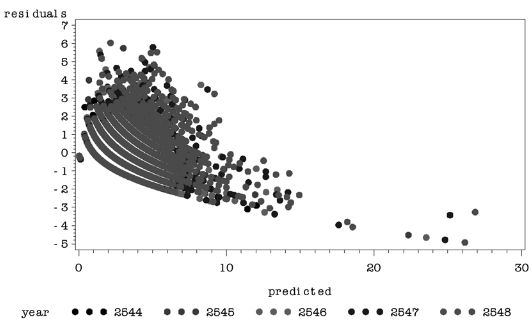
R_p^2 คือ ค่าร้อยละของการทำนายถูกต้อง

และจากตารางที่ 3 พบว่า การวิเคราะห์ด้วยตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) พบว่า ปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 เหมือนกัน คือ ค่าสินไหมทดแทนที่แปลงค่า อายุของรถยนต์ กลุ่มอายุผู้ทำประกันภัย อาชีพ และเพศ ส่วนการวิเคราะห์ตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ CS มีปัจจัยเช่นเดียวกันกับโครงสร้างความแปรปรวนร่วมแบบ AR(1) เมื่อตรวจสอบความเหมาะสมของตัวแบบพบว่า ค่าสถิติของการวิเคราะห์ตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) และ CS มีค่าสถิติ Generalized Chi-Square/DF เท่ากับ 0.18 และ 0.15 ตามลำดับ ซึ่งมีค่าน้อยกว่า 1 แสดงว่าไม่เกิดปัญหาได้ค่าความแปรปรวนมากเกินจริง และพิจารณาค่าร้อยละของความถูกต้องที่ได้จากการทำนายของตัวแบบแสดงภาพรวมของค่าเฉลี่ยประชากรมีค่าเท่ากับ 53.34 และ 53.38 ตามลำดับ ส่วนค่าร้อยละของความถูกต้องที่ได้จากการทำนายของตัวแบบเมื่อเพิ่มเทอมอิทธิพลของค่าคงที่สุ่มในแต่ละบุคคลหรือหน่วยศึกษามีค่าเท่ากับ 76.28 และ 79.36

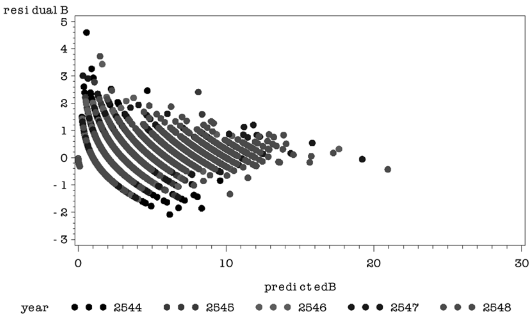
3. ผลการตรววจสอบความเหมาะสมของตัวแบบที่พัฒนาจากตัวแบบ GEE และตัวแบบ GLMMs ว่ามีความเหมาะสมกับข้อมูลหรือไม่ โดยทำแผนภาพการกระจายของค่า Residuals คู่กับ Predicted หรือค่าประมาณบนเส้นถดถอย ในระยะเวลาติดตาม 5 ปี ของการทำประกันภัยรถยนต์พบว่า การกระจายของข้อมูลมีความสัมพันธ์กันลักษณะแบบเส้นโค้ง หรืออยู่ในรูปแบบเอกซ์โปเนนเชียล ซึ่งตรงตามกับลักษณะของตัวแบบที่ตัวแปรอยู่ในรูปของล็อกการิทึม โดยตัวแปรตามหรือจำนวนการเรียกค่าสินไหมทดแทนมีการแจกแจงแบบปัวส์ซง ดังนั้นตัวแบบจึงมีความเหมาะสมกับข้อมูลที่ใช้ในการศึกษา ดังแสดงในรูปที่ 1 ถึง 4



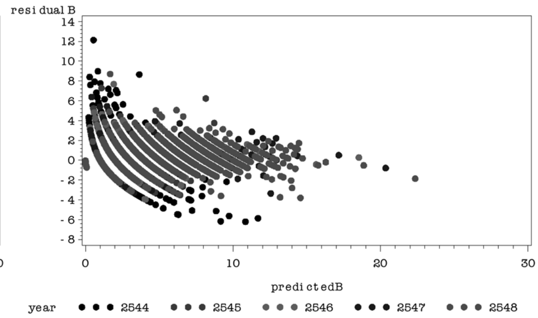
รูปที่ 1 การพล็อตค่า Residuals คู่กับ Predicted ($\hat{Y}$) ของตัวแบบ GEE เมื่อมีรูปแบบ AR(1)



รูปที่ 2 การพล็อตค่า Residuals คู่กับ Predicted ($\hat{Y}$) ของตัวแบบ GEE เมื่อมีรูปแบบ CS



รูปที่ 3 การพล็อตค่า Residuals คู่กับ Predicted $(\hat{Y}_u)$ ของตัวแบบ GLMMs เมื่อมีรูปแบบ AR(1)



รูปที่ 4 การพล็อตค่า Residuals คู่กับ Predicted $(\hat{Y}_u)$ ของตัวแบบ GLMMs เมื่อมีรูปแบบ CS

กำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) มีความเหมาะสมสำหรับใช้เป็นตัวแบบของข้อมูลที่ใช้ศึกษานี้มากกว่าตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) จากค่า R_p^2 ของตัวแบบสำหรับภาพรวมของค่าเฉลี่ยประชากรเท่ากับร้อยละ 53.34 และ 52.32 ตามลำดับ โดยตัวแบบ GLMMs สามารถอธิบายการเปลี่ยนแปลงของจำนวนการเรียกค่าสินไหมทดแทนได้ทั้งภาพรวมค่าเฉลี่ยประชากรและแต่ละหน่วยศึกษาของผู้ทำประกันภัยรถยนต์สามารถเขียนสมการของตัวแบบที่เหมาะสมได้ คือ

4.1 ตัวแบบสำหรับภาพรวมของค่าเฉลี่ยประชากร

$\log_e y$ = -2.7130 + 0.1805 (ค่าสินไหมทดแทนที่แปลงค่า)* + 0.0861 (อายุรถยนต์)*
+ 0.1390 (กลุ่มอายุ 18-24 ปี) + 0.1703 (กลุ่มอายุ 25-35 ปี)* + 0.0744 (กลุ่มอายุ 36-50 ปี)*
+ 0.3105 (อาชีพ)* + (-0.0487) (เพศ)*

(* มีนัยสำคัญทางสถิติที่ระดับ 0.05)

4.2 ตัวแบบสำหรับแต่ละหน่วยศึกษาเมื่อมีเทอมอิทธิพลสุ่มของหน่วยศึกษา (Random Subject Effect)

$\log_e y$ = (-2.7130 + Random Subject Effect) + 0.1805 (ค่าสินไหมทดแทน)* + 0.0861 (อายุรถยนต์)*
+ 0.1390 (กลุ่มอายุ 18-24 ปี) + 0.1703 (กลุ่มอายุ 25-35 ปี)* + 0.0744 (กลุ่มอายุ 36-50 ปี)*
+ 0.3105 (อาชีพ)* + (-0.0487) (เพศ)*

(* มีนัยสำคัญทางสถิติที่ระดับ 0.05)

เช่น หน่วยศึกษาที่ 1 (Subject 1) มีค่าคงที่ = 0.4135 โดยที่ -2.2995 มาจาก -2.7130 + 0.4135 ดังนั้นตัวแบบสำหรับหน่วยศึกษาที่ 1 เขียนได้เป็น

$\log_e y$ = -2.2995 + 0.1805 (ค่าสินไหมทดแทนที่แปลงค่า)* + 0.0861 (อายุรถยนต์)* + 0.1390 (อายุผู้ทำประกันภัย 18-24 ปี) + 0.1703 (อายุผู้ทำประกันภัย 25-35 ปี)*
+ 0.0744 (อายุผู้ทำประกันภัย 36-50 ปี)* + 0.3105 (อาชีพ)* + (-0.0487) (เพศ)*

การอธิบายความหมายของตัวแบบ GLMMs เมื่อโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) บางตัวแปรสำหรับข้อมูลทั้งหมด คือ

สำหรับตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ

ถ้าปัจจัยอายุรถยนต์เพิ่มขึ้น 1 ปี ทำให้จำนวนการเรียกค่าสินไหมทดแทนเพิ่มขึ้น

โดยเฉลี่ยเป็น 1.09 ครั้ง/คน/ปี เมื่อปัจจัยอื่นๆ ในตัวแบบคงที่ [exp(0.0861) = 1.09]

สำหรับตัวแปรอิสระเป็นข้อมูลเชิงคุณภาพ

กลุ่มอายุ 18-24 ปี จะมีจำนวนการเรียกค่าสินไหมทดแทนโดยเฉลี่ยปีน 1.15 ครั้ง/คน/ปี เมื่อเทียบกับกลุ่มอายุเกิน 50 ปี โดยที่ปัจจัยอื่นๆ ในตัวแบบคงที่ [exp(0.1390) = 1.15]

กลุ่มอายุ 25-35 ปี จะมีจำนวนการเรียกค่าสินไหมทดแทนโดยเฉลี่ยเป็น 1.19 ครั้ง/คน/ปี เมื่อเทียบกับกลุ่มอายุเกิน 50 ปี โดยที่ปัจจัยอื่นๆ ในตัวแบบคงที่ [exp(0.1703) = 1.19]

กลุ่มอายุ 36-50 ปี จะมีจำนวนการเรียกค่าสินไหมทดแทนโดยเฉลี่ยเป็น 1.08 ครั้ง/คน/ปี เมื่อเทียบกับกลุ่มอายุเกิน 50 ปี โดยที่ปัจจัยอื่นๆ ในตัวแบบคงที่ [exp(0.0744) = 1.08]

ผู้ทำประกันภัยรถยนต์ที่เป็นเพศชายจะมีจำนวนการเรียกค่าสินไหมทดแทนโดยเฉลี่ยเป็น 1.36 ครั้ง/คน/ปี เมื่อเทียบกับผู้ทำประกันภัยรถยนต์ที่เป็นเพศหญิง โดยที่ปัจจัยอื่นๆ ในตัวแบบคงที่ [exp(0.3105) = 1.36]

สรุปและวิจารณ์ผลการศึกษา

ผลการศึกษาสรุปได้ว่า ตัวแบบ GEE เมื่อกำหนดโครงสร้างความสัมพันธ์ของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) เป็นตัวแบบที่มีความเหมาะสมกับข้อมูล โดยมีปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 คือ ค่าสินไหมทดแทนที่แปลงค่า อายุรถยนต์ อายุผู้ทำประกันภัย 25-35 ปี อายุผู้ทำประกันภัย 36-50 ปี และอาชีพ ส่วนตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) มีความเหมาะสมกับข้อมูลเช่นเดียวกันมีปัจจัยที่มีผลต่อจำนวนการเรียกค่าสินไหมทดแทนที่ระดับนัยสำคัญทางสถิติ 0.05 คือ ค่าสินไหมทดแทนที่แปลงค่า อายุรถยนต์ อายุผู้ทำประกันภัย 25-35 ปี อายุผู้ทำประกันภัย 36-50 ปี อาชีพและเพศ เมื่อทำการเปรียบเทียบตัวแบบทั้งสองวิธีสรุปได้ว่า ตัวแบบ GLMMs เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของจำนวนการเรียกค่าสินไหมทดแทนเป็นแบบ AR(1) มีความเหมาะสมสำหรับใช้เป็นตัวแบบของข้อมูลจำนวนการเรียกค่าสินไหมทดแทนในการศึกษานี้ทั้งตัวแบบแสดงภาพรวมค่าเฉลี่ยประชากรและตัวแบบสำหรับแต่ละหน่วยศึกษาเมื่อมีเทอมอิทธิพลค่าคงที่สุ่ม

สำหรับผลที่ได้จากการศึกษานี้มีค่าร้อยละของการทำนายถูกต้องไม่สูงมากนัก ผู้วิจัยจึงเห็นว่าหากมีการเพิ่มตัวแปรอื่นที่เกี่ยวข้องกับจำนวนการเรียกค่าสินไหมทดแทน เช่น พฤติกรรมการใช้รถยนต์ ประเภทของการใช้รถยนต์ และคะแนนในการขับขี่ [4] เป็นต้น ที่มีความสัมพันธ์กับจำนวนการเรียกค่าสินไหมทดแทนจะทำให้ตัวแบบมีความเหมาะสมกับข้อมูลและมีประสิทธิภาพในการทำนายสูงขึ้นได้

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณกรมการประกันภัยที่ได้อนุเคราะห์ข้อมูลการประกันภัยรถยนต์สำหรับประกอบการทำการวิจัยนี้ให้สำเร็จลุล่วงด้วยดี

เอกสารอ้างอิง

1. Denuit, M. 1997. A New Distribution of Poisson Type for the Number of Claims. *ASTN Bulletin* 27: 229-242.
2. Stroinski, K. J., and Currie, I. D. 1989. Selection of Variables for Automobile Insurance Rating. *Insurance: Mathematics and Economics* 8(1): 35-46.
3. Liang, K. Y., and Zeger, S. L. 1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73: 13-22.
4. Yau, K. K. W., Yip, K. C. H., and Yuen, H. K. 2003. Modelling Repeated Insurance Claim Frequency Data Using the Generalized Linear Mixed Model. *Journal of Applied Statistic* 30: 857-865.
5. ประทีป ประดิษฐ์วรคุณ. 2540. การวิเคราะห์เชิงสถิติเพื่อหาอัตราเบี้ยประกันภัยรถยนต์ที่เหมาะสม. วิทยานิพนธ์มหาบัณฑิต พาณิชยศาสตร์และการบัญชี (การประกันภัย). กรุงเทพฯ. บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
6. กรมการประกันภัย. 2550. ประเภทของการประกันภัยรถยนต์. ได้จาก http://www.doi.go.th. 1 พฤษภาคม 2550.
7. กันทิมา ศิริพาณิชย์. 2545. การวิเคราะห์ปัจจัยที่มีผลต่อความเสียหายเพื่อการกำหนดอัตราดอกเบี้ยประกันภัยรถยนต์ในประเทศไทย. วิทยานิพนธ์มหาบัณฑิต สถิติ (การประกันภัย) จุฬาลงกรณ์มหาวิทยาลัย.
8. กัลยา วานิชย์บัญชา. 2549. การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล. ครั้งที่ 9. กรุงเทพฯ. บริษัทธรรมสาร จำกัด.
9. Jong, P., and Heller, G. Z. 2008. Generalized Linear Models for Insurance Data. 1st Edition. United Kingdom. Cambridge University Press.
10. Kutner, M. H., Nachtsheim, C. J., and Neter, J. 2004. Applied Linear Regression Models. 4th Edition. New York. The McGraw-Hill Companies, Inc.
11. Hosmer, D. W., and Lemeshow, S. 2000. Applied Logistic Regression. United Kingdom. Wiley Chichester.

12. Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. 2004. Applied Longitudinal Analysis. New Jersey. Wiley.
13. Gordon, J., and Maura, S. 1997 Applications of GEE Methodology Using the SAS System. SAS Institute Inc. Cary, NC.
14. Oliver, S. 2006. Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models. SAS Institute Inc. Cary, NC.
15. Wolfinger, R., and O'Connell, M. 1993. Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation* 48: 233-243.
16. Shoukri, M. M., and Chaudhary, M. A. 2007. Analysis of Correlated Data with SAS and R. 3rd Edition. USA. Taylor & Francis Group, LLC.

ได้รับบทความวันที่ 27 มกราคม 2552
ยอมรับตีพิมพ์วันที่ 12 มีนาคม 2552