SPU 53rd ANNIVERSARY

NATIONAL AND INTERNATIONAL SRIPATUM UNIVERSITY CONFERENCE 2023

The 18th National and The 8th International Sripatum University Conference

27 OCTOBER 2023

“หนังสือประมวลบทความ PROCEEDINGS”

การประชุมวิชาการระดับชาติ ครั้งที่ 18
และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8
มหาวิทยาลัยศรีปทุม

เรื่อง “วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน”

Organized by Sripatum University in cooperation with
• The University of Palermo, Italy • Sholokhov Moscow State University for the Humanities,Russia
• Universidad de Colima, Mexico • University of Taipei, Taiwan
• Institut Teknologi Sepuluh Nopember, Indonesia • The Joint Graduate School of Energy and Environment
• The Social Science Research Association of Thailand • Lawyers Council Under the Royal Patronage
• Thai Federation on Logistics • The Institute of Internal Auditors of Thailand • Prachachuen Research Network
• Journal Network of Social Sciences and Humanities • Program Management Unit for Human Resources
& Institutional Development, Research and Innovation (PMU-B)

หนังสือประมวลบทความ (Proceedings)
การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการ ระดับนานาชาติ ครั้งที่ 8
มหาวิทยาลัยศรีปทุม ออนไลน์ ประจำปี 2566
เรื่อง วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน
(Research and Innovations to Sustainable Development)

วันศุกร์ที่ 27 ตุลาคม 2566

รวบรวมโดย

คณะกรรมการพิจารณาผลงาน

การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8 ประจำปี 2566

ออกแบบปกโดย งานกราฟิกและศิลปกรรม มหาวิทยาลัยศรีปทุม

จัดรูปเล่มโดย โรงพิมพ์ มหาวิทยาลัยศรีปทุม

หนังสือประมวลบทความ (Proceedings)

การประชุมวิชาการระดับชาติ ครั้งที่ 18 และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 8 มหาวิทยาลัยศรีปทุม ออนไลน์

เรื่อง วิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน

The Proceedings of the 18th National and the 8th International Sripatum University Conference : Research and Innovations to Sustainable Development

วันที่: 27 ตุลาคม 2566

Date: 27 October 2023

ISBN (e-book): 978-974-655-469-5

ข้อมูลทางบรรณานุกรมของหอสมุดแห่งชาติ

หนังสือประมวลบทความการประชุมวิชาการระดับชาติ ครั้งที่ 18 และระดับนานาชาติ ครั้งที่ 8 มหาวิทยาลัยศรีปทุม ออนไลน์ เรื่อง การวิจัยและนวัตกรรมสู่การพัฒนาที่ยั่งยืน.-- พิมพ์ครั้งที่ 18.-- กรุงเทพฯ: มหาวิทยาลัยศรีปทุม, 2566.

3383 หน้า.

1. การประชุมวิชาการ. 2. บทความวิจัย. 3. บทความวิชาการ. I. ชื่อเรื่อง.

060

เจ้าของ

มหาวิทยาลัยศรีปทุม

จัดทำโดย

ศูนย์ส่งเสริมการวิจัยและการประกันคุณภาพการศึกษา มหาวิทยาลัยศรีปทุม

สถานที่จัดพิมพ์และจัดทำรูปเล่ม

โรงพิมพ์ มหาวิทยาลัยศรีปทุม

2410/2 ถนนพหลโยธิน แขวงเสนานิคม เขตจตุจักร กรุงเทพฯ 10900 โทร. 02 579 1111 ต่อ 1552

สารบัญ

สารบัญบทความ (ต่อ)

หน้า

การวิเคราะห์ของตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้น สำหรับข้อมูลข้อเรียกร้องประกันภัยดูแลสุขภาพ

The Analysis of the Double Generalized Linear Model for Health Care Insurance Claims Data

สุจินดา พินิจชัย
สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี
E-mail: sujinda2540@gmail.com
ณฐกร นวรตน
สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี
E-mail: natakonnawaratana@gmail.com
อมรรัตน์ สุริยวิจิตรเศรณี
สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี
E-mail: amornrat@g.sut.ac.th
เจษฎา ตัณฑนุช
สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี
E-mail: jessada@g.sut.ac.th

บทคัดย่อ

ตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นเป็นตัวแบบที่พัฒนาต่อยอดมาจากตัวแบบเชิงเส้นวางนัยทั่วไป เพื่อให้สามารถใช้ในการพยากรณ์ได้อย่างมีประสิทธิภาพมากขึ้น งานวิจัยนี้ได้วิเคราะห์และเปรียบเทียบตัวแบบทั้งสองเพื่อใช้ในการพยากรณ์ค่าใช้จ่ายในการรักษาที่บริษัทประกันจ่ายให้กับผู้ถือกรรมธรรม์ ข้อมูลที่ใช้ในการทำวิจัยเป็น “Sample Insurance Claim Prediction Dataset” ของ Eason ปรับปรุงล่าสุดเมื่อวันที่ 4 มิถุนายน 2018 การพยากรณ์ได้ใช้วิธีสร้างตัวแบบเชิงเส้นทั้งสองชนิดจากข้อมูลของผู้ถือกรมธรรม์ได้แก่ อายุ เพศ ดัชนีมวลกาย จำนวนก้าวเฉลี่ยที่เดินใน 1 วัน จำนวนบุตร การเป็นผู้สูบบุหรี่หรือไม่ ภูมิภาคที่อยู่ (ในประเทศอเมริกา) และการเป็นผู้เรียกร้องค่าสินไหมประกันหรือไม่ โดยพิจารณาว่าข้อมูลมีการแจกแจงทางสถิติที่เป็นไปได้ 4 ประเภท ได้แก่ ปรกติ แกมมา อินเวอร์สเกาส์เซียน และทวีดี การสร้างตัวแบบใช้วิธีแบ่งข้อมูลออกเป็นข้อมูลสำหรับการเรียนรู้และการทดสอบในอัตราส่วน 70:30, 75:25 และ 80:20 ผลการศึกษาพบว่าตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นที่มีสมมติฐานว่าข้อมูลมีการกระจายทางสถิติแบบทวีดีให้ประสิทธิภาพในการพยากรณ์ดีที่สุด เมื่อใช้สร้างตัวแบบด้วยอัตราส่วนการเรียนรู้และการทดสอบ 75:25 โดยมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 5345.859

คำสำคัญ: ตัวแบบเชิงเส้นวางนัยทั่วไป, ตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้น, การแจกแจงทางสถิติ, ประกันภัยดูแลสุขภาพ

ABSTRACT

A double generalized linear model is an advanced model developed from the generalized linear model to improve forecasting efficiency. This research analyzed and compared both models for predicting healthcare expenses paid by insurance companies to policyholders. The data used in this study is the 'Sample Insurance Claim Prediction Dataset' by Eason, last updated on June 4, 2018. The prediction utilized the two models created from data of policyholders including age, gender, body mass index, average walking steps per day, the number of children, smoking status, the region of policyholder in the United States, and whether the claimant has health insurance or not. It was assumed that the data follows one of four statistical distributions: normal, gamma, inverse-Gaussian, and Tweedie. Model development involved splitting the data into training and testing sets in various ratios, such as 70:30, 75:25, and 80:20. The study found that the double generalized linear model, assuming that the data follows a Tweedie distribution, performed the best in terms of prediction accuracy when trained and tested with a 75:25 split, yielding an average root mean square error (RMSE) of 5345.859.

Keywords: generalized linear model, double generalized linear model, statistical distribution, health care insurance

1. ความสำคัญและที่มาของปัญหาวิจัย

ปัจจุบันมีการพัฒนาทางด้านระบบสารสนเทศและเทคโนโลยีดิจิทัลอย่างก้าวกระโดด ทำให้เกิดการประยุกต์ใช้คณิตศาสตร์และสถิติในการวิเคราะห์ข้อมูลที่มีความซับซ้อนเพิ่มขึ้นเป็นอย่างมาก องค์ความรู้ดังกล่าวช่วยให้สามารถทำความเข้าใจสถานการณ์และสามารถพยากรณ์คาดการณ์เหตุการณ์ต่าง ๆ ได้อย่างมีประสิทธิภาพ ไม่ว่าจะเป็นการใช้ในการพยากรณ์สภาพอากาศและวิเคราะห์ปริมาณผลผลิตทางการเกษตร (Sankar et al., 2019) การวางแผนการจัดสรรด้านการเงินและการลงทุน รวมไปถึงการพยากรณ์โรคระบาด ประมวลผลข้อมูลทางการแพทย์ และการประกันสุขภาพ (ณฐกร นวรตน และคณะ, 2566) คณิตศาสตร์และสถิติช่วยให้เราสามารถสร้างตัวแบบที่สามารถใช้ในการวิเคราะห์ข้อมูลและพยากรณ์ได้อย่างน่าเชื่อถือและแม่นยำ ตัวแบบที่เป็นที่นิยมมากในการใช้พยากรณ์ข้อมูลต่าง ๆ คือ **ตัวแบบเชิงเส้นวางนัยทั่วไป** (generalized linear model) หรือ GLM ตัวแบบชนิดนี้มีข้อดีหลากหลาย เช่น ใช้งานได้ง่าย เข้าใจได้ง่าย สามารถปรับใช้กับข้อมูลได้หลากหลาย และที่สำคัญคือสามารถอธิบายความสัมพันธ์ระหว่าง**ตัวแปรอธิบาย** (explanatory variables) และ**ตัวแปรตอบสนอง** (response variable) ได้ อย่างไรก็ตามตัวแบบ GLM ก็มีข้อเสียในแง่ความซับซ้อนในการปรับแต่งตัวแบบ มีความไวต่อข้อมูลที่มีความผิดปกติ รวมไปถึงสมมติฐานบางอย่างของตัวแบบไม่ตรงกับความเป็นจริงของข้อมูล (Frees, 2010) ในการสร้างตัวแบบ GLM จะเริ่มด้วยสมมติฐานว่าพารามิเตอร์การกระจาย (dispersion parameter) จะต้องถูกกำหนดไว้ก่อน ขณะที่ความเป็นจริงของข้อมูลอาจมีปัจจัยอื่นภายนอกที่ทำให้พารามิเตอร์ดังกล่าวไม่เป็นไปตามที่ตั้งไว้ ในปี 1989 Smyth ได้นำเสนอตัวแบบ GLM ที่มีการกระจายแปรผันได้ (Generalized Linear Models with Varying Dispersion) ซึ่งภายหลังมักถูกเรียกในชื่อ**ตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้น** (double generalized linear model) หรือ DGLM (Smyth, 1989) ซึ่งตัวแบบชนิดนี้ต่างจาก GLM ในแง่ที่ต้องสร้างตัวแบบที่มีทั้งตัวแบบค่าเฉลี่ย (mean model) และตัวแบบการกระจาย (dispersion model) ไปพร้อม ๆ กัน (Nelder & Lee, 1998) ทำให้ DGLM มีประสิทธิภาพดีกว่า GLM ถ้าหากว่าข้อมูลที่ต้องการพยากรณ์มีความสัมพันธ์กับตัวแปรอธิบายในแบบ

ไม่เชิงเส้น (nonlinear relation) และ มีความไวต่อข้อมูลที่มีความผิดปกติน้อยกว่า GLM และสามารถอธิบายความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตอบสนองได้ดีกว่า GLM

จากแนวคิดข้างต้นพบว่า DGLM เป็นการสร้างตัวแบบที่น่าสนใจมาก แต่ยังไม่พบว่ามีงานวิจัยในประเทศไทยที่มีการประยุกต์ใช้ DGLM ในการพยากรณ์และการศึกษาในเชิงลึก งานวิจัยชิ้นนี้ต้องการศึกษาถึง DGLM เพื่อเป็นต้นแบบในการทำวิจัยในด้านอื่น ๆ รวมทั้งตัวอย่างที่ประยุกต์ใช้ในข้อมูลข้อเรียกร้องประกันภัยดูแลสุขภาพ โดยจะเปรียบเทียบประสิทธิภาพกับ GLM ทั้งนี้การทำตัวแบบทั้งหมดจะทำภายใต้สมมติฐานว่าข้อมูลอาจมีการแจกแจงทางสถิติ 4 ประเภท ได้แก่ การแจกแจงปรกติ การแจกแจงแกมมา การแจกแจงอินเวอร์สเกาส์เซียน และ การแจกแจงทวีดี

2. วัตถุประสงค์ของการวิจัย

เพื่อศึกษา วิเคราะห์และเปรียบเทียบประสิทธิภาพของตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นกับตัวแบบเชิงเส้นวางนัยทั่วไปโดยพิจารณาจากข้อมูลข้อเรียกร้องประกันภัยดูแลสุขภาพ

3. เอกสารและงานวิจัยที่เกี่ยวข้อง

3.1 การแจกแจงทางสถิติที่เกี่ยวข้องกับงานวิจัย

การแจกแจงปรกติ (Normal Distribution)

การแจกแจงปรกติหรือบางครั้งอาจรู้จักในอีกชื่อว่าการแจกแจงเกาส์เซียน (Gaussian distribution) เป็นฟังก์ชันทางสถิติที่เป็นที่นิยมในการอธิบายความน่าจะเป็นของการเกิดผลลัพธ์ที่เป็นไปได้ต่าง ๆ ของเหตุการณ์ที่เราสนใจเมื่อพิจารณาว่าตัวแปรสุ่ม (random variable) ของเหตุการณ์นั้นมีความต่อเนื่อง (continuous) การแจกแจงนี้มีฟังก์ชันความหนาแน่นของความน่าจะเป็น (probability density function-PDF) ที่มีลักษณะเป็นเส้นโค้งรูประฆัง (bell curve) ในงานด้านประกันภัยการแจกแจงปรกติมักถูกใช้เพื่อประมาณการความน่าจะเป็นของการเรียกร้องค่าสินไหมทดแทน วิเคราะห์ความเสี่ยง ประมาณความสูญเสียที่อาจจะเกิดขึ้นในกรณีที่เกิดเหตุการณ์ไม่คาดคิด ซึ่งนำไปสู่การคาดการณ์เพื่อกำหนดเบี้ยประกันที่เหมาะสม อย่างไรก็ตามการใช้การแจกแจงปรกติในงานด้านประกันภัยอาจมีข้อจำกัดเนื่องด้วยสมมติฐานว่า PDF มีลักษณะสมมาตรซ้าย-ขวา แต่บางครั้งความเสี่ยงงานที่เกี่ยวข้องกับประกันภัยอาจมีการกระจายที่ไม่เป็นไปตามสมมติฐานดังกล่าว (Leinwander & Aziz, 2018; Nawatana, 2019)

การแจกแจงแกมมา (Gamma Distribution)

การแจกแจงแกมมาเป็นการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่มักใช้ในการจำลองตัวแปรที่มีลักษณะการสะสมหรืออัตราการเกิดที่เพิ่มขึ้นตามเวลา มักใช้ในการจำลองระยะเวลาระหว่างสองเหตุการณ์ การแจกแจงแบบแกมมามี 2 พารามิเตอร์ได้แก่ พารามิเตอร์กำหนดรูปร่าง (shape parameter) α และพารามิเตอร์กำหนดมาตรา (scale parameter) β ทั้งนี้หาก α มีค่าน้อยการแจกแจงแกมมาจะมีความเบ้ (skewness) ไปทางขวาและในทางกลับกันหาก α มีค่ามากการแจกแจงจะมีความเบ้ไปทางซ้าย การแจกแจงแกมมาจะเป็นการแจกแจงแบบเลขชี้กำลัง (exponential distribution) เมื่อ $\alpha=1$ และเป็นการแจกแจงปัวซง (Poisson distribution) เมื่อ α เข้าสู่ค่าอนันต์ สำหรับพารามิเตอร์กำหนดมาตรา หาก β มีค่าน้อย การแจกแจงจะมีความกระจายกว้างและหาก β มีค่ามาก การแจกแจงจะมีความกระจายแคบ การแจกแจงแกมมาเป็นฟังก์ชันความน่าจะเป็นที่อธิบายการกระจายของข้อมูลที่มีความแปรปรวนสูง ในงานด้านประกันภัยการแจกแจงแกมมามักถูกใช้เพื่อประมาณการความน่าจะเป็น

ของเหตุการณ์ต่าง ๆ เช่น การเรียกร้องค่าสินไหมทดแทนที่สูงผิดปกติ การประมาณการจำนวนผู้ป่วยที่อาจเข้ารับการรักษาในโรงพยาบาลด้วยโรคร้ายแรงในแต่ละปี การประมาณการจำนวนความเสียหายจากภัยธรรมชาติที่เกิดขึ้นในแต่ละปี การประมาณการจำนวนลูกค้าที่จะยกเลิกการประกันภัยในแต่ละปี (Nova, 2023)

การแจกแจงอินเวอร์สเกาส์เซียน (Inverse-Gaussian Distribution)

การแจกแจงชนิดนี้มีอีกชื่อเรียกว่าการแจกแจงวาล์ด (Wald distribution) การแจกแจงอินเวอร์สเกาส์เซียนเป็นการแจกแจงความน่าจะเป็นแบบต่อเนื่องที่มีเซตค้ำจุน (support) เป็นจำนวนจริงบวกเหมือนกับการแจกแจงแกมมา โดยมักถูกใช้ในการจำลองระยะเวลาที่ใช้ในสถานการณ์บางอย่างที่ถูกกำหนดไว้ เช่น เวลาที่ลูกค้ามาถึงสถานที่ให้บริการ ระยะเวลาที่อุปกรณ์หนึ่งทำงานจนชำรุด หรือระยะเวลาที่ผู้ป่วยจะเสียชีวิต นอกจากนี้ยังมีการใช้การแจกแจงดังกล่าวในการสร้างตัวแบบผลตอบแทนรายวันของหุ้น (daily stock returns) และความผันผวนของราคาสัญญาซื้อขายล่วงหน้า (uncertainty of future stock price) การแจกแจงอินเวอร์สเกาส์เซียนมี 2 พารามิเตอร์ได้แก่ พารามิเตอร์ค่าเฉลี่ย (mean parameter) μ และพารามิเตอร์กำหนดมาตรา (scale parameter) λ การแจกแจงนี้มีลักษณะเบ้ขวา พารามิเตอร์กำหนดมาตราจะเป็นตัวกำหนดความเบ้ของการแจกแจก หากค่าพารามิเตอร์สูงขึ้นการแจกแจงจะเบ้ไปทางซ้ายมากขึ้น การแจกแจงอินเวอร์สเกาส์เซียนถูกนำมาใช้ในด้านประกันภัยในการสร้างตัวแบบสำหรับการเรียกร้องค่าสินไหมทดแทนการบาดเจ็บทางร่างกาย (bodily injury claim) (Punzo, 2019)

การแจกแจงทวิดี (Tweedie Distribution)

การแจกแจงทวิดีถูกนำเสนอขึ้นครั้งแรกในงานประชุมวิชาการ the Indian Statistical Golden Jubilee International Conference ปี 1984 โดยมีลักษณะพิเศษคือ สามารถกลายเป็นการแจกแจงทางสถิติอื่น ๆ ที่เป็นที่รู้จักดีหากได้มีการปรับเปลี่ยนพารามิเตอร์ โดยการแจกแจงทวิดีมี 3 พารามิเตอร์ได้แก่ พารามิเตอร์ค่าเฉลี่ย μ พารามิเตอร์การกระจาย σ^2 และพารามิเตอร์กำลัง (power parameter) ρ ซึ่งเป็นพารามิเตอร์ที่กำหนดรูปร่างของการแจกแจงทวิดี ทั้งนี้ การแจกแจงปรกติคือกรณีพิเศษของการแจกแจงทวิดีเมื่อ ρ=0 การแจกแจงป๊วซงคือกรณีพิเศษของการแจกแจงทวิดีเมื่อ ρ=1 การแจกแจงแกมมาคือกรณีพิเศษของการแจกแจงทวิดีเมื่อ ρ=2 และการแจกแจงอินเวอร์สเกาส์เซียนคือกรณีพิเศษของการแจกแจงทวิดีเมื่อ ρ=3 (Dunn & Smyth, 2005) การแจกแจงทวิดีถูกนำมาใช้ในการสร้างตัวแบบหลายชนิด เช่น การสร้างตัวแบบของระดับมลพิษ (pollution levels) การสร้างตัวแบบทางด้านเศรษฐศาสตร์ การสร้างตัวแบบผลตอบแทนด้านการเงิน (financial returns) รวมไปถึง การสร้างตัวแบบการเรียกร้องค่าสินไหมประกัน (insurance claim) และการแจกแจงทวิดีถูกนำไปใช้อย่างมากในการสร้างตัวแบบวางนัยทั่วไป (Frees, 2010)

การแจกแจงทางสถิติที่กล่าวมาข้างต้นเป็นการแจกแจงวงศ์เลขชี้กำลัง (Exponential family distribution) ที่มีฟังก์ชันความน่าจะเป็นในรูป $f(x) = c(x,\phi)\exp\left(\frac{x\theta - a(\theta)}{\phi}\right)$ เมื่อ x คือตัวแปรสุ่ม สัญกรณ์ θ และ ϕ เป็นพารามิเตอร์ โดยเรียก θ ว่า **พารามิเตอร์แบบบัญญัติ** (canonical parameter) และ เรียก ϕ ว่า **พารามิเตอร์การกระจาย** (dispersion parameter) ซึ่งฟังก์ชัน $a(\theta)$ และ $c(x,\phi)$ มีค่าแตกต่างกันขึ้นกับการแจกแจงทางสถิตินั้น ๆ เช่น การแจกแจงปรกติ $\theta = \mu$ เป็นค่าเฉลี่ย $\phi = \sigma^2$ เป็นความแปรปรวน $c(x,\phi) = \frac{1}{\sqrt{2\pi\phi}}\exp(\frac{-x^2}{\phi})$ และ $a(\theta) = \frac{\theta^2}{2}$

3.2 ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function)

ฟังก์ชันภาวะน่าจะเป็นแสดงถึงความน่าจะเป็นของพารามิเตอร์ในตัวแบบทางสถิติโดยพิจารณาจากข้อมูลที่มี เช่น ตัวแปรสุ่ม X ที่มีการแจกแจงความน่าจะเป็นตามฟังก์ชันความน่าจะเป็น $f(x;\theta)$ สำหรับ

พารามิเตอร์ θ เป็นการพิจารณาความน่าจะเป็นของการเกิดเหตุการณ์ x ในทางกลับกันฟังก์ชันภาวะน่าจะเป็นของพารามิเตอร์ θ ณ ค่าตรึงของ x (fixed x) แทนด้วย $L(\theta; x)$ หรือ $L(\theta)$ เป็นค่าที่แสดงถึงความเป็นไปได้ของ θ สำหรับการกำหนดเหตุการณ์ x และในกรณีการเกิดเหตุการณ์ร่วม n เหตุการณ์ ค่าของ $L(\theta; x_1, x_2, \dots, x_n)$ จะมีค่าเท่ากับฟังก์ชันความน่าจะเป็นร่วม $f(x_1, x_2, \dots, x_n; \theta)$ ของตัวแปรสุ่ม $X_1, \dots, X_n$ เพื่อให้สะดวกต่อการคำนวณมักใช้การเปลี่ยนรูปด้วยฟังก์ชันลอการิทึมเพื่อแปลงฟังก์ชันความน่าจะเป็นให้อยู่ในรูปผลรวมเชิงเส้น ได้ผลลัพธ์เป็น**ฟังก์ชันล็อกภาวะน่าจะเป็น** (loglikelihood function) และในงานวิจัยนี้ฟังก์ชันล็อกภาวะน่าจะเป็นสำหรับการแจกแจงวงศ์เลขชี้กำลังของตัวแปรสุ่ม $X_1, \dots, X_n$ มีค่าดังนี้

$$\ell(\phi, \theta) = \sum_{i=1}^{n} \left\{ \ln c(x_i, \phi) + \frac{x_i\theta - a(\theta)}{\phi} \right\}$$

แม้ว่าฟังก์ชันภาวะน่าจะเป็นมีความใกล้เคียงกับ PDF แต่ทั้งสองฟังก์ชันมีการนำใช้งานและให้ความหมายที่แตกต่างกัน ซึ่ง PDF ใช้ในการคำนวณความน่าจะเป็นของการเกิดค่าใด ๆ ในการแจกแจงของตัวแปรสุ่ม ส่วนฟังก์ชันภาวะน่าจะเป็นบอกถึงความเป็นไปได้ของพารามิเตอร์ จึงนิยมใช้ในการประมาณค่าพารามิเตอร์ในตัวแบบโดยเลือกค่าพารามิเตอร์ที่ทำให้ค่าฟังก์ชันภาวะน่าจะเป็นมีค่าสูงที่สุด และเรียกวิธีดังกล่าวว่า วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood method)

3.3 ตัวแบบวางนัยทั่วไป (Generalized Linear Model - GLM)

GLM เป็นตัวแบบทางสถิติที่ใช้ในการวิเคราะห์ข้อมูลเพื่อช่วยในการคาดการณ์ ที่มีความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอธิบาย โดยองค์ประกอบของ GLM ประกอบด้วย 3 ส่วนหลัก ๆ ดังนี้

3.3.1 องค์ประกอบเชิงสุ่ม (random component) เป็นคุณสมบัติของการแจกแจงทางสถิติของตัวแปรตอบสนองซึ่งมีการระบุการแจกแจงอยู่ในวงศ์เลขชี้กำลัง (exponential family) ที่เหมาะสม

3.3.2 องค์ประกอบเชิงระบบ (systematic component) เป็นสมการเชิงเส้นที่อธิบายความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอธิบาย ได้แก่ $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in}$ ถูกเรียกว่า**ตัวพยากรณ์เชิงเส้น** (linear predictor) โดยที่ β_j คือพารามิเตอร์ลำดับที่ j ของตัวแปรอธิบาย $X_{ij}, j = 1, \dots, n$

3.3.3 ฟังก์ชันเชื่อมโยง (link function) ทำหน้าที่แปลงค่าคาดหมายของตัวแปรตอบสนองให้เป็นตัวพยากรณ์เชิงเส้น $g(\theta_i) = \eta_i$ โดยการใช้ฟังก์ชันเชื่อมโยงต้องเป็นฟังก์ชันแบบทางเดียว (monotonic) และสามารถหาอนุพันธ์ได้ ซึ่งจะมีลักษณะต่างกันขึ้นกับการกระจายตัวของข้อมูล เช่น ฟังก์ชันเชื่อมโยงเอกลักษณ์ คือ $g(\theta) = \theta$ หรือ ฟังก์ชันเชื่อมโยงลอการิทึม คือ $g(\theta) = \ln\theta$

ในการสร้างตัวแบบ GLM มีสมมติฐานที่สำคัญ คือ ความแปรปรวนของค่าคาดหวังของตัวแปรตอบสนองต้องเป็นค่าคงตัวกับทุกระดับของตัวแปรอธิบาย (Paula, 2013)

3.4 ตัวแบบวางนัยทั่วไปสองชั้น (Double Generalized Linear Model - DGLM)

สมมติฐานในการสร้างตัวแบบ GLM ที่ความแปรปรวนเป็นค่าคงตัวนั้น ไม่สอดคล้องกับข้อมูลในบางสถานการณ์ที่ความแปรปรวนของตัวแปรอธิบายมีความแตกต่างกันหรือข้อมูลที่ไม่มีความสมดุล ตัวแบบ DGLM จึงถูกพัฒนาขึ้นเพื่อใช้กับลักษณะดังกล่าว ซึ่งตัวแบบ DGLM เป็นการขยายแนวคิดจากตัวแบบ GLM ที่มีองค์ประกอบหลัก 3 ส่วนเหมือนกับ GLM แต่ในองค์ประกอบเชิงระบบ จะมีตัวพยากรณ์เชิงเส้น 2 ตัวได้แก่ ตัวพยากรณ์ของพารามิเตอร์แบบบัญญัติและตัวพยากรณ์ของพารามิเตอร์การกระจายตัว ตามลำดับดังนี้

$g(\theta_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in}$ และ $h(\phi_i) = \lambda_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_n Z_{in}$ เมื่อ β_0 และ γ_0 เป็นพารามิเตอร์ระยะตัดแกน (intercept parameter) สำหรับ β_j และ γ_j เป็นพารามิเตอร์ของตัวแปรอธิบาย X_{ij} และ Z_{ij} $(j = 1, \dots, n)$ ตามลำดับ โดยมี $g(\cdot)$ และ $h(\cdot)$ เป็นฟังก์ชัน

เชื่อมโยง การประมาณค่าพารามิเตอร์อาจจะใช้วิธีภาวะน่าจะเป็นสูงสุดหรือวิธีการประมาณค่าแบบอื่น เช่น ภาวะน่าจะเป็นสูงสุดกำกัด (Restricted Maximum Likelihood-REML) ทั้งนี้DGLM ยังใช้ตัวแบบการกระจายเลขชี้กำลังสองชั้น (Double Exponential Dispersion Mode-DEDM) ในการปรับค่าคาดหวังและค่าระยะตัดแกนให้เหมาะกับข้อมูล

3.5 การประเมินประสิทธิภาพ

ในงานวิจัยนี้ใช้เครื่องมือเพื่อประเมินประสิทธิภาพตัวแบบ ด้วยการวัดค่าความคลาดเคลื่อน 2 ชนิด ได้แก่ **ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย** (mean absolute error-MAE) และ**รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย** (root mean square error-RMSE)

ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย หรือ MAE คำนวณโดย $\mathrm{MAE} = \frac{1}{N}\sum|\mathrm{Error}|$ เมื่อ N คือจำนวนข้อมูล เป็นการหาค่าเฉลี่ยของขนาดความคลาดเคลื่อน (error magnitude) ในการพยากรณ์ของตัวแบบ ยิ่งมีค่าน้อยแสดงว่าตัวแบบสามารถพยากรณ์ได้แม่นยำ MAE มักถูกใช้งานในงานที่ต้องการการประเมินความคล้ายคลึง (similarity) หรือขนาดความคลาดเคลื่อนในการพยากรณ์แต่ละข้อมูลมีค่าไม่แตกต่างกัน

รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย หรือ RMSE คำนวณโดย $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum(\mathrm{Error})^2}$ แสดงให้เห็นถึงค่าเฉลี่ยของขนาดความคลาดเคลื่อนกำลังสอง ยิ่งมีค่าน้อยแสดงว่าตัวแบบสามารถพยากรณ์ได้แม่นยำเช่นกัน แต่ RMSE ยังแสดงให้เห็นถึงการกระจายตัวของขนาดความคลาดเคลื่อนด้วย หากตัวแบบมีค่า MAE ที่เท่ากันแต่มีค่า RMSE ที่น้อยกว่า แสดงให้เห็นว่าขนาดของความคลาดเคลื่อนมีการกระจายตัวที่น้อยกว่า ตัวแบบที่ขนาดของความคลาดเคลื่อนมีการกระจายตัวที่ผิดปรกติมักมีค่า RMSE ที่มาก

3.6 สมมติฐานการวิจัย

ตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นสำหรับการพยากรณ์ค่าใช้จ่ายในการรักษาที่บริษัทประกันจ่ายให้กับผู้เรียกร้องค่าสินไหมทดแทนมีประสิทธิภาพมากกว่าตัวแบบเชิงเส้นวางนัยทั่วไป

4. วิธีดำเนินการวิจัย

4.1 ประชากรและตัวอย่าง

ในงานวิจัยนี้ข้อมูลที่ใช้ในการดำเนินการได้ถูกรวบรวมจากฐานข้อมูลสาธารณะที่เรียกว่า "Sample Insurance Claim Prediction Dataset" ซึ่งถูกพัฒนาขึ้นบนฐานข้อมูล "Medical Cost Personal Datasets" ซึ่งเสนอโดย Eason ฐานข้อมูลนี้ถูกสร้างขึ้นเมื่อวันที่ 14 พฤษภาคม 2018 และได้รับการปรับปรุงล่าสุดเมื่อวันที่ 4 มิถุนายน 2018 สามารถเข้าถึงฐานข้อมูลนี้ได้ที่ https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset ฐานข้อมูลดังกล่าวอยู่ในรูปแบบไฟล์ CSV (comma-separated values file) ซึ่งประกอบด้วยรายการข้อมูลการเรียกเก็บค่าประกันภัยทั้งหมด 1,338 รายการที่รวบรวมข้อมูลจากผู้ถือกรมธรรม์ที่อาศัยในสหรัฐอเมริกา แต่ละรายการประกอบด้วยตัวแปรทั้งหมด 9 อย่างที่เกี่ยวข้องกับผู้ถือกรมธรรม์ (policyholder) ได้แก่ 1) Y=charges (ค่าใช้จ่ายที่บริษัทจ่ายให้กับผู้เรียกร้องค่าสินไหมทดแทน) 2) X_1=age (อายุ) 3) X_2=sex (เพศ) 4) X_3=BMI (ดัชนีมวลกาย) 5) X_4=steps (จำนวนก้าวที่เดินเฉลี่ยใน 1 วัน) 6) X_5=children (จำนวนบุตร) 7) X_6=smoker (เป็นผู้สูบบุหรี่หรือไม่) 8) X_7=region (ภูมิภาคที่อยู่ในประเทศอเมริกา) และ 9) X_8= insurance claim (การเป็นผู้เรียกร้องค่าสินไหมประกันหรือไม่)

4.2 เครื่องมือวิจัย

โปรแกรม RStudio ถูกใช้เป็นซอฟต์แวร์หลักในการสร้างตัวแบบและดำเนินการวิเคราะห์ทางสถิติในงานวิจัยนี้ เนื่องจากมีชุดคำสั่งทางสถิติและได้รับการยอมรับอย่างแพร่หลาย งานวิจัยนี้ได้ใช้งาน RStudio รุ่น 3.6.1 ทำงานภายใต้ระบบปฏิบัติการ Microsoft Windows 10 Pro รุ่น 21H2 การคำนวณถูกดำเนินการบนคอมพิวเตอร์ที่มีหน่วยประมวลผล (CPU) Intel I5-6200U หน่วยความจำ (RAM) 12GB

4.3 ขั้นตอนการดำเนินการวิจัย

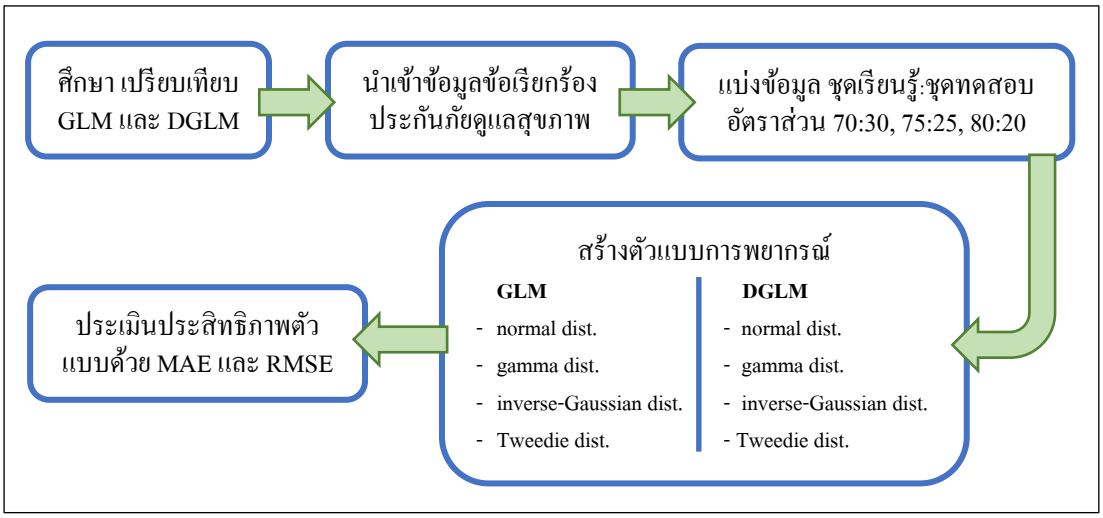
4.3.1 นำเข้าข้อมูลจากแฟ้มข้อมูลประเภท CSV ด้วยโปรแกรม RStudio และกำหนดตัวแปรดังที่กล่าวมาในหัวข้อ 4.2

4.3.2 แบ่งข้อมูลเป็นสองชุดได้แก่ ข้อมูลชุดเรียนรู้สำหรับสร้างตัวแบบและข้อมูลชุดทดสอบประสิทธิภาพผลการพยากรณ์ โดยแบ่งเป็นสัดส่วน 70:30 (937 ชุดข้อมูลสำหรับเรียนรู้และ 401 ชุดข้อมูลสำหรับทดสอบ) 75:25 (1,004 ชุดข้อมูลสำหรับเรียนรู้และ 334 ชุดข้อมูลสำหรับทดสอบ) และ 80:20 (1,071 ชุดข้อมูลสำหรับเรียนรู้และ 267 ชุดข้อมูลสำหรับทดสอบ)

4.3.3 สร้างตัวแบบการพยากรณ์ค่าใช้จ่ายในการรักษาของผู้เรียกร้องค่าสินไหมทดแทน (charges) โดยใช้ตัวแบบเชิงเส้นวางนัยทั่วไปที่ขึ้นอยู่กับผลรวมเชิงเส้นของตัวแปรอธิบายภายใต้การแจกแจง 4 แบบ ได้แก่ ปรกติ แกมมา อินเวอร์สเกาส์เซียน และทวีดี รวมถึงสร้างตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นที่มีตัวพยากรณ์ของพารามิเตอร์แบบบัญญัติและตัวพยากรณ์ของพารามิเตอร์การกระจายตัวขึ้นอยู่กับผลรวมเชิงเส้นของตัวแปรอธิบายภายใต้การแจกแจงทั้งสี่ดังกล่าว

4.3.4 ประเมินประสิทธิภาพการพยากรณ์ของตัวแบบโดยใช้ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย MAE และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย RMSE

4.4 แบบแผนการวิจัย



แผนภาพที่ 1 ขั้นตอนในการดำเนินการวิจัย

5. ผลการวิจัย

ผลการสร้างตัวแบบ GLM และตัวแบบ DGLM โดยพิจารณาว่าข้อมูลมีการแจกแจงทางสถิติที่เป็นไปได้ 4 แบบ และแบ่งเป็นข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบเป็นอัตราส่วนต่าง ๆ ทั้ง 3 อัตราส่วน แสดงประสิทธิภาพด้วยการวัดค่าคลาดเคลื่อน MAE และ RMSE แสดงดังตารางที่ 1

ตารางที่ 1 แสดงประสิทธิภาพของตัวแบบ GLM และ DGLM แบบต่าง ๆ

การแจกแจงความน่าจะเป็น		อัตราส่วนในการแบ่งเป็นข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบ					
		70:30		75:25		80:20	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
ปรกติ	GLM	3862.959	5895.714	3735.222	5381.199	4042.047	6105.660
	DGLM	3856.384	5888.988	3791.467	5354.916	3916.951	6009.750
แกมมา	GLM	5108.133	10680.460	5032.179	9061.112	4797.476	8412.457
	DGLM	4091.081	5731.783	3990.248	5377.105	4325.139	6001.553
อินเวอร์สเกาส์เซียน	GLM	7044.871	16024.340	5683.105	11291.860	7524.664	18107.800
	DGLM	4108.400	5868.850	3987.566	5425.047	4308.082	6073.463
ทวิดี	GLM	**3848.274**	6384.640	**3671.664**	6422.211	**3868.104**	6597.916
	DGLM	4126.012	**5605.711**	4036.015	**5345.859**	4341.212	**5897.221**

จากตารางที่ 1 แสดงให้เห็นว่าหากพิจารณาว่าข้อมูลมีการแจกแจงทางสถิติแบบทวิดีจะให้ประสิทธิภาพในการสร้างตัวแบบทั้ง GLM และ DGLM ที่ดีกว่า ไม่ว่าจะแบ่งเป็นข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบอัตราส่วนใดก็ตาม โดยหากพิจารณาเฉพาะความคลาดเคลื่อน MAE ตัวแบบ GLM ที่ข้อมูลถูกแบ่งด้วยอัตราส่วน 75:25 มีประสิทธิภาพสูงที่สุด โดยตัวพยากรณ์เชิงเส้นมีค่าดังนี้

$$g(E(Y)) = 7.335+2.527\times10^{-2}X_1-5.766\times10^{-2}X_2+2.309\times10^{-2}X_3+4.778\times10^{-6}X_4+4.351\times10^{-2}X_5+1.414X_6-4.063\times10^{-2}X_7-1.005\times10^{-1}X_8$$

เมื่อ $E(Y)$ คือ ค่าคาดหมาย (expected valued) และหากพิจารณาเฉพาะค่าคลาดเคลื่อน RMSE ตัวแบบ DGLM ที่ข้อมูลถูกแบ่งด้วยอัตราส่วน 75:25 มีประสิทธิภาพสูงที่สุดเช่นกัน โดยตัวพยากรณ์ของพารามิเตอร์แบบบัญญัติมีค่า

$$g(E(Y)) = 7.573+1.093\times10^{-2}X_1-8.476\times10^{-3}X_2+3.565\times10^{-2}X_3-2.592\times10^{-5}X_4+2.726\times10^{-2}X_5+1.247X_6-1.156\times10^{-2}X_7+9.680\times10^{-2}X_8$$

และตัวพยากรณ์ของพารามิเตอร์การกระจายมีค่า

$$h(\phi) = 3.718-2.986\times10^{-2}X_1+3.746\times10^{-2}X_2+1.155\times10^{-2}X_3-2.874\times10^{-6}X_4-2.229\times10^{-1}X_5-1.695X_6-4.120\times10^{-2}X_7-2.678\times10^{-1}X_8$$

6. อภิปรายผล

ผลการวิจัยแสดงให้เห็นว่าการสมมติให้ข้อมูลข้อเรียกร้องประกันภัยดูแลสุขภาพมีการแจกแจงทางสถิติแบบทวิดีทำให้การสร้างตัวแบบเพื่อพยากรณ์มีประสิทธิภาพสูงกว่าการสมมติว่าข้อมูลมีการแจกแจงทางสถิติแบบอื่นที่เหลืออีก 3 แบบ ซึ่งสอดคล้องกับที่ปรากฏใน (Goldburd et. al., 2020) ทั้งนี้เพราะว่าการแจกแจงปรกติ แกมมา และอินเวอร์เกาส์เซียน เป็นกรณีพิเศษของการแจกแจงทวิดี และการแบ่งข้อมูลเพื่อให้เหมาะสมต่อการสร้างตัวแบบพยากรณ์และใช้ในการทดสอบประสิทธิภาพตัวแบบ คืออัตราส่วน 75:25 ทั้งนี้เมื่อพิจารณาเปรียบเทียบการสร้างตัวแบบพยากรณ์ด้วย GLM และ DGLM เนื่องด้วยการใช้เครื่องมือวัดประสิทธิภาพที่ต่างกันทำให้เห็นว่า ได้ตัวแบบที่มีประสิทธิภาพสูงสุดที่แตกต่างกัน หากพิจารณาเฉพาะความคลาดเคลื่อนสัมบูรณ์เฉลี่ย MAE ในการสร้างตัวแบบด้วย GLM การดำเนินการจะกำหนดค่าพารามิเตอร์การกระจายไว้ก่อน แล้วทำการสร้างตัวพยากรณ์เชิงเส้นที่ให้ค่าคลาดเคลื่อนน้อยที่สุด ทำให้พบว่าตัวแบบ GLM ที่ได้มีค่า MAE ต่ำที่สุด อย่างไรก็ตามเมื่อพิจารณาค่าคลาดเคลื่อน RMSE พบว่าตัวแบบชนิดนี้มีค่า RMSE สูงกว่าตัวแบบที่ถูกสร้างด้วย DGLM แสดงให้เห็นว่าค่าคลาดเคลื่อนของตัวแบบ GLM นี้ มีการกระจายตัวที่มากกว่า ในทางกลับกันในการสร้างตัวแบบ

DGLM จะต้องสร้างตัวพยากรณ์เชิงเส้น 2 ตัวได้แก่ ตัวพยากรณ์ของพารามิเตอร์แบบบัญญัติและตัวพยากรณ์ของพารามิเตอร์การกระจายตัว การสร้างตัวแบบดังกล่าวจำเป็นต้องสมดุลระหว่างค่าคลาดเคลื่อนที่น้อยที่สุดของทั้งสองตัวพยากรณ์ ส่งผลให้ตัวแบบที่ได้จาก DGLM อาจจะไม่ได้มีค่า MAE ที่ต่ำที่สุด แต่ก็ยังมีค่า MAE ที่ต่ำรวมถึงมีค่า RMSE ที่ต่ำที่สุด แสดงว่าตัวแบบ DGLM ที่สร้างขึ้นมีค่าคลาดเคลื่อนในการพยากรณ์ของตัวแบบที่การกระจายตัวต่ำกว่าตัวแบบ GLM ดังนั้นจะถือว่าตัวแบบพยากรณ์ที่ถูกสร้างด้วย DGLM เป็นตัวแบบที่มีประสิทธิภาพที่สูงกว่า

7. ข้อเสนอแนะ

7.1 ข้อเสนอแนะในการนำผลวิจัยไปใช้

ผลการวิจัยแสดงให้เห็นว่าตัวแบบเชิงเส้นวางนัยทั่วไปสองชั้นมีประสิทธิภาพในการพยากรณ์ที่ดีและยืดหยุ่นกว่าตัวแบบเชิงเส้นวางนัยทั่วไป ควรถูกนำไปประยุกต์ใช้ในการพยากรณ์ด้านอื่น ๆ นอกเหนือจากงานทางด้านประกันภัย เช่น การวางแผนการจัดสรรด้านการเงิน การจัดพอร์ตการลงทุน

7.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

ในการประเมินประสิทธิภาพของตัวแบบอาจใช้เครื่องมืออื่นในการประเมินเช่น เกณฑ์สารสนเทศของอะกะกิเกะ หรือเอไอซี (Akaike's information criterion-AIC) เกณฑ์สารสนเทศของเบส์ หรือบีไอซี (Bayesian information criterion-BIC) และ Distance between Indices of Simulation and Observation (DISO)

8. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนโดยทุนการศึกษากิตติบัณฑิตของมหาวิทยาลัยเทคโนโลยีสุรนารี (มทส.) โครงการพัฒนาและส่งเสริมผู้มีความสามารถทางด้านวิทยาศาสตร์และเทคโนโลยี (พสวท.) และสาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

9. เอกสารอ้างอิง

ณฐกร นวรตน, พันงาม วงศ์คำจันทร์, สุขเกษม วัชรมัยสกุล และ เจษฎา ตัณฑนุช (2566), การประยุกต์ใช้วิศวกรรมคุณลักษณะและตัวแบบเชิงเส้นนัยทั่วไปสำหรับพยากรณ์จำนวนผู้ติดเชื้อใหม่ไวรัสโคโรนา 2019, *วารสารวิทยาศาสตร์บูรพา*, ปีที่ 28 (ฉบับที่ 1) มกราคม – เมษายน พ.ศ. 2566, หน้า 552-569.

Dunn, P. and G. Smyth (2005). Series evaluation of Tweedie exponential dispersion models. *Statistics and Computing. 15*, 267–280.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2020). *Generalized Linear Models for Insurance Rating (2nd ed.)*. Casualty Actuarial Society. https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf

Leinwander, A. J., & Aziz, M. A. (2018). Modeling Insurance Claims Using Flexible Skewed and Mixture Probability Distributions. *Journal of Modern Applied Statistical Methods, 17*(1), eP2467. doi: 10.22237/jmasm/1525133100

Nawatana, N. (2019). *Analysis of distributions for insurance claims data.* Master of Science in Applied Mathematics thesis, Suranaree University of Technology.

Nelder, J. A., & Lee, Y. (1998). Joint modeling of mean and dispersion. *Technometrics, 40*(2), 168-171.

Nova. (2023, March 27). Gamma Distribution Explained: Applications in Finance, Engineering, and More. *AI Tech Trend*. https://aitechtrend.com/gamma-distribution-explained-applications-in-finance-engineering-and-more/#insurance-claims

Paula, G. A. (2013). On diagnostics in double generalized linear models. *Computational Statistics & Data Analysis, 68*, 44–51, DOI: 10.1016/j.csda.2013.06.008

Punzo, A. (2019) A new look at the inverse Gaussian distribution with applications to insurance and economic data, *Journal of Applied Statistics, 46*(7), 1260-1287, DOI: 10.1080/02664763.2018.1542668

Sankar, T., Prasad, A., Sudarmanian, N.S., & Ganesan, D. (2019). Weather Forecasting in Agriculture. In *Research Trends in Agriculture Sciences*, pp. 57-81). AkiNik Publications.

Smith, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society - Series B, 51*, 47-60.