# ความครอบคลุมและยืดหยุ่น : ประเด็นที่ควรถูกพิจารณาสำหรับการวิเคราะห์ข้อมูลด้วยตัวแบบเชิงเส้นนัยทั่วไปในงานวิจัยทางวิทยาศาสตร์สุขภาพ

พงษ์เดช สารการ [1, 2], ฎลกร จำปาหวาย [2]

(วันที่รับบทความ:30 กรกฎาคม 2563; วันที่แก้ไข 30 ตุลาคม 2563; วันที่ตอบรับ 31 ตุลาคม 2563)

**บทคัดย่อ**

ตัวแบบเชิงเส้นนัยทั่วไป (Generalized linear model, GLMs) เป็นตัวแบบที่ขยายมาจากตัวแบบเชิงเส้นทั่วไป (General linear model, GLM) เพื่อพัฒนาสมการทำนาย หรือ ความสัมพันธ์เชิงเส้นระหว่างตัวแปรผลลัพธ์และตัวแปรร่วม ซึ่งครอบคลุมทั้งตัวแปรผลลัพธ์แบบต่อเนื่องและไม่ต่อเนื่องที่อยู่ภายใต้ฟังก์ชั่นการแจกแจงตระกูลเอกซโพเนนเชียล ด้วยส่วนประกอบเชิงสุ่มและฟังก์ชั่นเชื่อมโยง นอกจากนี้ยังถูกพัฒนาและขยายอย่างต่อเนื่องเป็นตัวแบบและวิธีการอื่นที่สามารถนำไปใช้ในงานวิจัยที่มีความซับซ้อนด้วยเช่นกัน เช่น ตัวแบบ Generalized additive model (GAM) สำหรับความสัมพันธ์เชิงเส้นโค้งแบบราบเรียบและสมการประมาณค่านัยทั่วไป (Generalized estimating equation, GEE) สำหรับตัวแปรผลลัพธ์ที่มีความสัมพันธ์กัน เป็นต้น ดังนั้นประเด็นความครอบคลุมและยืดหยุ่นของตัวแบบเชิงเส้นนัยทั่วไป จึงควรถูกพิจารณาและนำมาใช้ในขั้นตอนการวิเคราะห์ข้อมูล เพื่อลดเวลาในการเรียนรู้วิธีการทางสถิติแต่ละวิธีและสะดวกในการนำมาใช้ โดยเฉพาะในงานวิจัยทางวิทยาศาสตร์สุขภาพ

**คำสำคัญ:** ตัวแบบเชิงเส้นทั่วไป, ตัวแบบเชิงเส้นนัยทั่วไป, ฟังก์ชั่นเชื่อมโยง

[1] รองศาสตราจารย์, สาขาวิชาวิทยาการระบาดและชีวสถิติ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น, E-mail: spongd@kku.ac.th

[2] ASEAN Cancer Epidemiology and Prevention Research Group (ACEP), คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น, E-mail: donlawj@kkumail.com

Corresponding Author: พงษ์เดช สารการ, E-mail: spongd@kku.ac.th

# Coverage and flexibility: Issues Should be Considered for Analyzing by Generalized Linear Model in Health Science Research

Pongdech Sarakarn [1, 2] , Donlagon Jumparway [2]

(Receive 30th July 2020; Revised: 30th October 2020; Accepted 31st October 2020)

**Abstract**

Generalized linear model (GLMs) is the model which extends from the general linear model (GLM) for developing the predictive equations or linear relationship between outcome and covariates, which covers both continuous and discrete outcomes based on the distribution of exponential family by random component and link function. Furthermore, such model is continuously developed and extended as various models and methods which can be used in the complicated research as well, such as the generalized additive model (GAM) for smoothing relationship and generalized estimating equation (GEE) for correlated outcomes. Therefore, flexibility and coverage issues of the generalized linear model should be considered and brought to use in the process of data analysis for saving the time-consuming of learning on each statistical method of statistics and being convenient for using, especially in health science research.

**Keywords:** general linear model, generalized linear model, link function

[1] Associate professor Dr., Department of Epidemiology and Biostatistics, Faculty of Public Health, Khon Kaen University, E-mail: spongd@kku.ac.th

[2] ASEAN Cancer Epidemiology and Prevention Research Group (ACEP), Faculty of Public Health, Khon Kaen University, E-mail: donlawj@kkumail.com

Corresponding Author: Pongdech Sarakarn, E-mail: spongd@kku.ac.th

**บทนำ**

ความครอบคลุม (Coverage) และการยืดหยุ่น (Flexibility) ถือเป็นประเด็นหนึ่งที่มีความสำคัญอย่างมากต่อการพัฒนา หรือ การปรับปรุงคุณภาพ เพราะประเด็นดังกล่าวสามารถสะท้อนให้เห็นถึงประสิทธิภาพของการพัฒนาในเรื่องนั้นได้เป็นอย่างดีว่า มีการพัฒนาอย่างต่อเนื่อง เพื่อลดข้อจำกัดเดิมที่มีอยู่ หรือ เพื่อสร้างแนวทางใหม่ ให้มีความยืดหยุ่น หรือ มีความครอบคลุมและสอดคล้องกับบริบทต่างๆ ได้มากขึ้นจากเดิม เช่นเดียวกับประเด็นวิธีการทางสถิติ (Statistical methods) ซึ่งถือเป็นเครื่องมือที่มีความสำคัญและจำเป็นอย่างมากในขั้นตอนการวิเคราะห์ข้อมูลสำหรับทุกงานวิจัย เพื่อให้ได้มาซึ่งข้อค้นพบที่มีความถูกต้องและแม่นยำ ภายใต้ความสอดคล้องกับข้อมูลที่รวบรวมมาได้ตามแบบแผนงานวิจัยและคำถามการวิจัยที่กำหนด โดยเฉพาะในสถานการณ์ปัจจุบัน ซึ่งปัญหาสุขภาพส่วนใหญ่เกิดขึ้นจากหลายปัจจัยที่สัมพันธ์และเกี่ยวข้องกันไปตามบริบทที่เปลี่ยนแปลงและแตกต่างกันตลอดเวลา จึงส่งผลให้ข้อมูลที่ได้และคำถามวิจัยที่เกี่ยวข้อง มีความซับซ้อนและหลากหลายมากขึ้น ดังนั้นวิธีการทางสถิติที่ถูกพัฒนาให้มีความครอบคลุมและยืดหยุ่นอย่างต่อเนื่อง เพื่อให้สอดรับกับสภาพปัญหาการวิจัยที่เปลี่ยนแปลงไป จึงเป็นประเด็นที่นักวิจัย โดยเฉพาะสาขาวิทยาศาสตร์สุขภาพ ควรให้ความสำคัญและนำมาพิจารณา เพื่อเป็นทางเลือกในการนำวิธีการทางสถิติที่เหมาะสม มาใช้วิเคราะห์ข้อมูลสำหรับงานวิจัยให้ผลลัพธ์ที่ได้มีความถูกต้องและแม่นยำมากยิ่งขึ้น

ตัวแบบเชิงเส้นนัยทั่วไป (Generalized linear model, GLMs) ถือเป็นตัวแบบทางสถิติที่ขยายมาจากตัวแบบเชิงเส้นทั่วไป (General linear model, GLM) สองคำนี้อาจดูใกล้เคียงและมีคำย่อคล้ายคลึงกัน โดยเฉพาะในบทความ หรือตำราบางเล่ม ดังนั้นเพื่อจำแนกความแตกต่างระหว่างสองชื่อได้ง่ายขึ้น นักวิจัยควรพิจารณาจากคำแรกของชื่อเรียก เช่น “**Generalized=นัยทั่วไป**” หรือ “**General=ทั่วไป**” เป็นต้น แต่เนื่องจากตัวแบบเชิงเส้นทั่วไป (GLM) เดิมพบว่า ยังมีข้อจำกัด ซึ่งถูกนำมาใช้วิเคราะห์ข้อมูลได้เพียงกับตัวแปรผลลัพธ์แบบต่อเนื่องเท่านั้น ภายใต้ข้อตกลงเบื้องต้นของความคลาดเคลื่อนที่ค่อนข้างเข้มงวด เช่น การแจกแจงแบบปกติ ความแปรปรวนคงที่และเท่ากัน (Homoscedasticity) และเป็นอิสระต่อกัน เป็นต้น [1] ซึ่งข้อตกลงเบื้องต้นดังกล่าว เมื่อนำมาใช้ในทางปฏิบัติมักพบว่า มีแนวโน้มการถูกละเมิดค่อนข้างสูง เนื่องจากไม่สอดคล้องกับบริบทของข้อมูลที่เป็นจริง โดยวิธีการทางสถิติที่นักวิจัยส่วนใหญ่คุ้นเคยและอยู่ภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) ดังกล่าว ได้แก่ สมการถดถอยเชิงเส้น การวิเคราะห์ความแปรปรวน (ANOVA) หรือการวิเคราะห์ความแปรปรวนร่วม (ANCOVA) เป็นต้น และจากข้อจำกัดดังกล่าว ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) จึงถูกพัฒนา เพื่อลดข้อจำกัดเดิมของชนิดข้อมูลตัวแปรผลลัพธ์ที่มีอยู่ให้สามารถนำมาใช้ได้ครอบคลุมมากขึ้น ทั้งข้อมูลแบบต่อเนื่องและไม่ต่อเนื่อง หรือ ข้อมูลที่มีการแจกแจงทั้งแบบปกติและไม่ปกติ นอกจากนี้ยังสามารถผ่อนคลายความเข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นของความคลาดเคลื่อนภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) เดิมที่มีอยู่ลงได้ ดังนั้นจึงทำให้นักวิจัยมีความสะดวกและยืดหยุ่นต่อการใช้งานภายใต้ตัวแบบนี้มากยิ่งขึ้น รวมถึงยังสามารถนำตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ไปขยายผลต่อ เป็นวิธีการประมาณค่า เพื่อวิเคราะห์ข้อมูลสำหรับคำถามวิจัยซึ่งตัวแปรผลลัพธ์สัมพันธ์กัน (Correlated outcome) เช่น สมการประมาณค่า

นัยทั่วไป (Generalized estimating equation, GEE) หรือ เป็นวิธีการทางสถิติ เพื่อวิเคราะห์ข้อมูลโดยคำนึงถึงผลกระทบในระดับเฉพาะบุคคล (Subject-specific effects) เช่น การวิเคราะห์ตัวแบบผลกระทบผสมเชิงเส้นนัยทั่วไป (Generalized linear mixed effects model, GLMM) เป็นต้น [2]

อย่างไรก็ตาม แม้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ถูกนำมาใช้อย่างแพร่หลายทั้งในต่างประเทศและบางสาขาของประเทศไทย แต่ในงานวิจัยทางวิทยาศาสตร์สุขภาพ ซึ่งข้อมูลตัวแปรผลลัพธ์ส่วนใหญ่ มีลักษณะแบบไม่ต่อเนื่อง หรือมีการแจกแจงแบบไม่ปกติ ได้แก่ ข้อมูลแบบแจงนับ (Categorical data) เช่น ผลลัพธ์แสดงอาการของโรค (ปกติ/ผิดปกติ) ผลลัพธ์แสดงความเสี่ยง (เสี่ยง/ไม่เสี่ยง) เป็นต้น หรือ ข้อมูลแบบจำนวนนับ (Count data) เช่น จำนวนครั้งของอุบัติเหตุที่เกิดขึ้นบนสี่แยกแห่งหนึ่งในรอบ 3 เดือนที่ผ่านมา เป็นต้น พบว่า มีการนำตัวแบบ GLMs ดังกล่าวมาใช้ในการวิเคราะห์ข้อมูลค่อนข้างน้อย แม้มีการนำมาใช้ในบางประเด็น เช่น วิธีการ GEE สำหรับการวิเคราะห์ข้อมูลกรณีตัวแปรผลลัพธ์สัมพันธ์กัน เป็นต้น แต่ยังพบว่า มีขอบเขตจำกัดเฉพาะกลุ่มหรือ สาขา ขณะเดียวกันจากการสืบค้นผ่านแหล่งเรียนรู้ หรือ แหล่งค้นคว้าอื่นบนระบบออนไลน์ทั้งในและต่างประเทศ ยังพบว่า ตำรา หรือ บทความวิชาการที่เผยแพร่เกี่ยวกับตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ส่วนใหญ่มีเนื้อหามุ่งเน้นการอธิบายในเชิงทฤษฎีภายใต้สัญลักษณ์ทางคณิตศาสตร์ที่ซับซ้อนเป็นหลัก มากกว่าการอธิบายในเชิงปฏิบัติ ด้วยการบ่งชี้ หรือ อธิบายถึงประเด็น หรือ การให้เหตุผลสำคัญที่สามารถนำมาพิจารณาและตัดสินใจเพื่อเลือกนำตัวแบบ GLMs ไปใช้แทนวิธีการทางสถิติเดิมที่มีอยู่อย่างชัดเจน จากประเด็นดังกล่าวจึงอาจเป็นข้อจำกัด หรือ อุปสรรคสำคัญต่อการเรียนรู้และการทำความเข้าใจ เพื่อนำตัวแบบ GLMs มาใช้ในการวิเคราะห์ข้อมูล โดยเฉพาะนักวิจัยที่ยังไม่คุ้นเคย หรือ ไม่ใช่นักสถิติ นอกจากนี้ยังอาจส่งผลกระทบต่อโอกาสในการพัฒนาผลงานวิจัยให้เป็นที่ยอมรับและสามารถตีพิมพ์เผยแพร่ในระดับที่สูงขึ้นได้

ดังนั้นบทความนี้ จึงมีวัตถุประสงค์เพื่อนำเสนอข้อมูลสำคัญเกี่ยวกับตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ให้นักวิจัยได้ทราบและเรียนรู้ด้วยตนเอง โดยมุ่งเน้นการอธิบายที่เป็นรูปธรรมและสามารถนำไปสู่การพิจารณาและตัดสินใจเลือกใช้ตัวแบบดังกล่าวได้ในทางปฏิบัติ ประกอบด้วยหัวข้อ ภาพรวมและข้อจำกัดของตัวแบบเชิงเส้นทั่วไป (GLM) ภาพรวมของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ความครอบคลุมและยืดหยุ่นของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) และข้อเสนอแนะการนำตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มาใช้ในงานวิจัยทางวิทยาศาสตร์สุขภาพ

**ภาพรวมและข้อจำกัดของตัวแบบเชิงเส้นทั่วไป (GLM)**

ในทางปฏิบัติ เมื่อกล่าวถึง “ตัวแบบ” หรือ “Model” สำหรับการวิเคราะห์ข้อมูลในงานวิจัย จะหมายถึง กรอบการทำงาน (Framework) แบบหนึ่งเดียว (Unity) ซึ่งมีความจำเพาะตามชื่อเรียกของตัวแบบ นั่นคือวิธีการทางสถิติใดที่อยู่ภายใต้ตัวแบบเดียวกัน ย่อมแสดงได้ว่า วิธีการทางสถิติดังกล่าว มีกรอบการทำงานเช่นเดียวกัน แม้ในทางปฏิบัติ วิธีการทางสถิติแต่ละวิธีดังกล่าว อาจถูกนำมาใช้กับคำถามวิจัย หรือ วัตถุประสงค์ที่แตกต่างกัน แต่ในทางทฤษฎี วิธีการทางสถิติเหล่านั้น ยังถือว่า อยู่ภายใต้กรอบทำงาน หรือ ตัวแบบเดียวกันเช่นเดิม

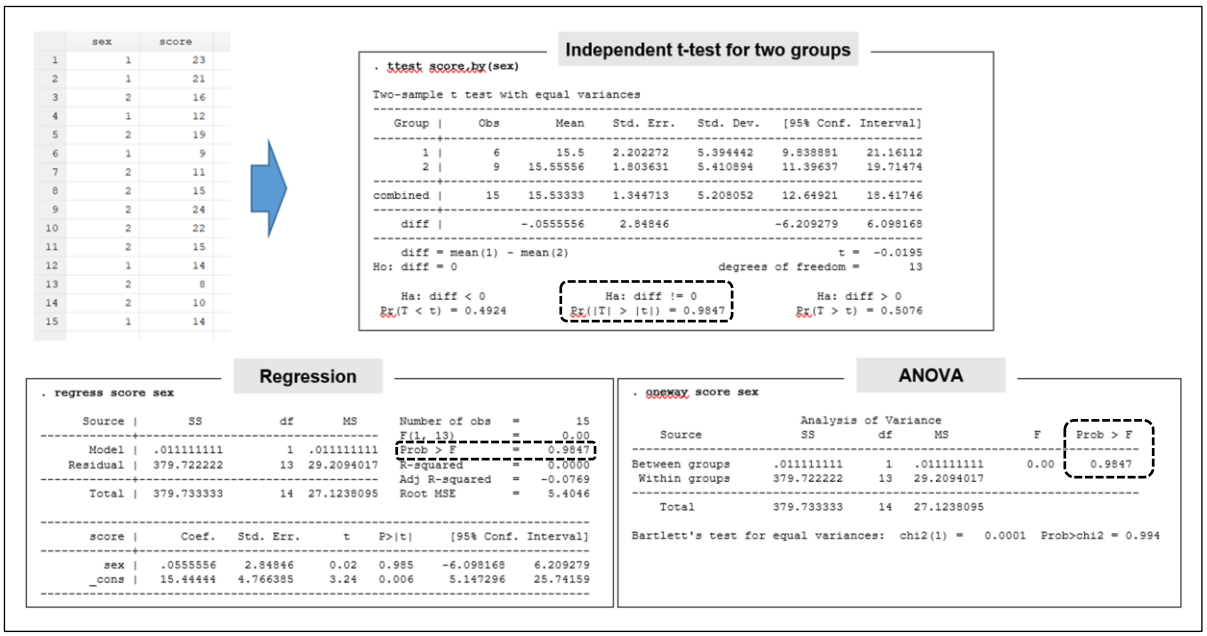
ตัวแบบเชิงเส้นทั่วไป (General linear model, GLM) ถือเป็นตัวแบบที่มีกรอบการทำงาน

หลักสำคัญ ประกอบด้วย ลักษณะข้อมูลตัวแปรผลลัพธ์ที่นำมาใช้ (เฉพาะตัวแปรผลลัพธ์แบบต่อเนื่องเท่านั้น) ข้อตกลงเบื้องต้นสำคัญของความคลาดเคลื่อน (ได้แก่ การแจกแจงแบบปกติ ความแปรปรวนคงที่และเท่ากัน ความสัมพันธ์เชิงเส้นตรงและความเป็นอิสระต่อกัน) วิธีการประมาณค่าพารามิเตอร์ (ใช้วิธีกำลังสองน้อยที่สุด หรือ Ordinary least square, OLS) และมีรูปสมการเชิงเส้นของตัวแบบเชิงเส้นทั่วไป (GLM) เป็นดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

เมื่อ $Y$ แทนตัวแปรผลลัพธ์แบบต่อเนื่อง, $\beta_0$ แทนค่าคงที่ หรือ จุดตัดบนแกน $Y$ (เมื่อกำหนดให้ $X$ ทุกตัวมีค่าเป็นศูนย์), $\beta_1,...,\beta_k$ แทนค่าสัมประสิทธิ์ถดถอยของตัวแปรร่วม ($X_1,...,X_k$) และ $\varepsilon$ แทนค่าความคลาดเคลื่อน

วิธีการทางสถิติที่นักวิจัยส่วนใหญ่คุ้นเคย เช่น สมการถดถอยแบบง่าย หรือ พหุคูณ การวิเคราะห์ t-test แบบอิสระสองกลุ่ม การวิเคราะห์ความแปรปรวน (ANOVA) และการวิเคราะห์ความแปรปรวนร่วม (ANCOVA) เป็นต้น ถือว่า ทั้งหมดอยู่ภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) และมีกรอบการทำงานแบบหนึ่งเดียวเช่นกัน นั่นคือ หากนักวิจัยต้องการนำวิธีการทางสถิติเหล่านี้ไปใช้งาน จำเป็นต้องพิจารณาและคำนึงถึงพื้นฐานการทำงาน ภายใต้กรอบแบบหนึ่งเดียวของตัวแบบเชิงเส้นทั่วไป (GLM) ซึ่งจุดเด่นของการมีกรอบทำงานเดียวกันดังกล่าว จะทำให้นักวิจัยมีความยืดหยุ่น โดยสามารถอธิบายความหมาย หรือ แปลผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล แม้เกิดขึ้นจากวิธีการวิเคราะห์ทางสถิติที่แตกต่างกัน ให้มีความเชื่อมโยงและสอดคล้องกัน ภายใต้ข้อสรุปที่เป็นไปในทิศทางเดียวกันได้ ดังเช่น นักวิจัยท่านหนึ่งสุ่มตัวอย่างนักเรียน จำนวน 15 คน มาทดสอบความรู้เรื่อง การแพร่ระบาดของโรค COVID-19 ในประเทศไทย (คะแนนเต็ม 30 คะแนน) เมื่อคำถามวิจัยต้องการทราบว่า “เพศชายและหญิงมีความรู้เรื่องการแพร่ระบาดฯ แตกต่างกัน หรือไม่ ? ” จากกรณีดังกล่าว นักวิจัยสามารถทำการวิเคราะห์ข้อมูลด้วยโปรแกรม STATA โดยพิจารณาเลือกวิธีการทางสถิติ 3 วิธี ที่อยู่ภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) เดียวกัน ได้แก่ การวิเคราะห์ t-test แบบอิสระสองกลุ่ม การวิเคราะห์สมการถดถอยแบบง่ายและการวิเคราะห์ความแปรปรวน มาทำการวิเคราะห์ เพื่อเปรียบเทียบผลลัพธ์ที่ได้ ซึ่งจากการวิเคราะห์ข้อมูล พบว่า ทั้งสามวิธีดังกล่าว ให้ผลลัพธ์ที่เหมือนกันและสามารถนำไปสู่การตัดสินใจและสรุปผลไปในทิศทางเดียวกันได้ เช่น กรณีการวิเคราะห์ t-test แบบสองกลุ่มอิสระและการวิเคราะห์ความแปรปรวน (ANOVA) สามารถสรุปผลเดียวกันได้ว่า “เพศชายและหญิง มีความรู้เรื่องการแพร่ระบาดฯ แตกต่างกัน อย่างไม่มีนัยสำคัญทางสถิติที่ระดับความผิดพลาดไม่เกิน 5%”, (p-value=0.9847) ขณะที่ในกรณีการวิเคราะห์สมการถดถอยแบบง่าย สามารถสรุปผลได้ว่า “เพศ มีผลต่อคะแนนความรู้เรื่องการแพร่ระบาดฯ อย่างไม่มีนัยสำคัญทางสถิติที่ระดับความผิดพลาดไม่เกิน 5%”, (p-value=0.9847) เป็นต้น ดังแผนภาพที่ 1




**แผนภาพที่ 1** แสดงผลการวิเคราะห์ข้อมูลด้วยวิธีการทางสถิติที่แตกต่างกัน ภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) เดียวกัน

อย่างไรก็ตาม แม้ตัวแบบเชิงเส้นทั่วไป (GLM) ได้ถูกนำมาใช้อย่างแพร่หลายและนักวิจัยส่วนใหญ่คุ้นเคยและมีประสบการณ์ทั้งการเรียนรู้และการนำมาใช้งานจริงในช่วงเวลาที่ผ่านมา แต่ในงานวิจัยบางสาขา โดยเฉพาะงานวิจัยทางวิทยาศาสตร์สุขภาพ ยังคงมีข้อจำกัดในการนำมาใช้ ดังรายละเอียดต่อไปนี้

**1.ตัวแบบเชิงเส้นทั่วไป (GLM) ถูกนำมาใช้ได้กับตัวแปรผลลัพธ์แบบต่อเนื่องเท่านั้น** จากข้อจำกัดนี้ จึงพบว่า ตัวแบบเชิงเส้นทั่วไป (GLM) ถูกนำมาใช้วิเคราะห์ข้อมูลในงานวิจัยทางวิทยาศาสตร์สุขภาพค่อนข้างน้อย เนื่องจากลักษณะข้อมูลในงานวิจัยสาขาดังกล่าว ส่วนใหญ่มีลักษณะแบบไม่ต่อเนื่อง หรือ มีการแจกแจงไม่ปกติ ซึ่งวิธีการทางสถิติภายใต้ตัวแบบ GLM ดังกล่าว ไม่สามารถรองรับได้

**2.ตัวแบบเชิงเส้นทั่วไป (GLM) มีข้อตกลงเบื้องต้นของความคลาดเคลื่อนค่อนข้างเข้มงวด โดยเฉพาะการแจกแจงแบบปกติและความเท่ากัน หรือ คงที่ของความแปรปรวน** ซึ่งข้อตกลงเบื้องต้นดังกล่าว สำหรับงานวิจัยทางวิทยาศาสตร์สุขภาพในทางปฏิบัติ พบว่า มีโอกาสเป็นไปได้น้อยมากที่จะสอดคล้องกับลักษณะข้อมูลจริงที่มีอยู่ เช่น ข้อมูลค่าใช้จ่ายที่เกิดขึ้นในระบบบริการสุขภาพของโรคต่างๆ ส่วนใหญ่มีลักษณะการแจกแจงแบบเบ้ขวา/บวก (Positively skewed distribution) หรือ มีการแจกแจงแบบไม่ปกติ เป็นต้น [3] และแม้ที่ผ่านมา ปัญหาการละเมิดข้อตกลงเบื้องต้นดังกล่าวของตัวแบบเชิงเส้นทั่วไป (GLM) ส่วนใหญ่ จะได้รับการแก้ไขและจัดการด้วยวิธีการแปลงค่าข้อมูล (Data transformation) โดยเฉพาะด้วยวิธี log transformation ซึ่งถูกนำมาใช้อย่างแพร่หลายและกว้างขวาง เพื่อจัดการกับข้อมูลที่มีลักษณะการแจกแจงแบบเบ้ขวา/บวก [4] แต่อย่างไรก็ตาม ยังพบว่า การแก้ไขปัญหาดังกล่าว ขึ้นกับเงื่อนไขหรือ คุณสมบัติที่เอื้อเฉพาะในบางกรณีเท่านั้น เช่น สามารถนำมาใช้ได้ในกรณีที่ข้อมูลมีการแจกแจงแบบ Log normal เดิมอยู่แล้ว [5] ขณะที่จากคุณสมบัติของ Log transformation ซึ่งไม่สามารถคำนวณค่าศูนย์ (0) และค่าติดลบ (-) ได้ เช่น

เมื่อคำนวณค่า log(0) หรือ ค่า log(-2) จะไม่สามารถคำนวณค่า Log ได้ เป็นต้น และหากข้อมูลที่นำมาวิเคราะห์มีค่าศูนย์ หรือ ค่าติดลบ นักวิจัยจะไม่สามารถนำวิธี Log transformation มาใช้ในการพิจารณาแปลงค่าข้อมูลทั้งหมดให้ครอบคลุมได้ ขณะเดียวกันเมื่อนักวิจัยแปลงค่าข้อมูลด้วยวิธี Log transformation แล้ว ในการสรุปผลท้ายสุด จำเป็นต้องแปลงค่าข้อมูลคืนกลับ (Back transformation) ซึ่งในขั้นตอนนี้อาจต้องคำถึงถึงความถูกต้องของสเกล (Scale) ที่เปลี่ยนไป รวมถึงเงื่อนไขทางคณิตศาสตร์ที่มีความซับซ้อนและยุ่งยาก และอาจนำไปสู่การสรุปผลที่ผิดพลาดได้ นอกจากนี้ยังมีประเด็นสำคัญ ซึ่งถูกกล่าวถึงในหลายการศึกษาที่ผ่านมาและให้ข้อสรุปไปในทิศทางเดียวกันว่า การแปลงข้อมูลด้วยวิธี Log transformation ภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) อาจไม่การันตีได้ว่า ปัญหาทั้งกรณีการแจกแจงแบบไม่ปกติและกรณีความแปรปรวนไม่เท่ากันจะถูกแก้ไขไปพร้อมกันทั้งสองกรณีในเวลาเดียวกันได้ เนื่องจากวิธีการ Log transformation มุ่งเน้นการทำให้ข้อมูลตัวแปรผลลัพธ์มีความสมมาตร (Symmetry) มากกว่าการแจกแจงแบบปกติ (Normal distribution) ขณะเดียวกันในทางปฏิบัติ การแปลงข้อมูลด้วย log transformation ส่วนใหญ่จะกระทำกับข้อมูลตัวแปรผลลัพธ์ภายใต้กลุ่มตัวอย่างที่นำมาศึกษา ซึ่งมีความผันแปรจากหลายปัจจัยที่แตกต่างกัน จึงทำให้การแปลงค่าข้อมูลที่ได้ขึ้นกับลักษณะความจำเพาะของแต่ละกรณีตัวอย่างที่นำมาศึกษา ดังนั้นการสรุปผลเพื่ออ้างอิงไปสู่ระดับประชากร อาจนำไปสู่ข้อสรุปที่คลาดเคลื่อน หรือ มีความลำเอียง โดยเฉพาะผลของการประมาณค่าสัมประสิทธิ์ถดถอยที่ได้ [3, 4, 6-8]

ซึ่งจากข้อจำกัดดังกล่าวข้างต้น ตัวแบบเชิงเส้นทั่วไป (GLM) จึงได้ถูกพัฒนาและกำหนดกรอบการทำงานแบบหนึ่งเดียวขึ้นมาใหม่ ด้วยการขยายผลต่อมาจากแนวคิดของตัวแบบ GLM เดิม เพื่อให้เกิดความครอบคลุมและมีความยืดหยุ่นในการนำไปใช้งานได้มากขึ้นและเรียกตัวแบบใหม่ ซึ่งพัฒนาและขยายผลมาจากตัวแบบเชิงเส้นทั่วไป (GLM) นี้ว่า “ตัวแบบเชิงเส้นนัยทั่วไป (Generalized linear model, GLMs)”

**ภาพรวมของตัวแบบเชิงเส้นนัยทั่วไป (GLMs)**

จากข้อจำกัดของตัวแปรผลลัพธ์ที่นำมาใช้และข้อตกลงเบื้องต้นที่เข้มงวดของตัวแบบเชิงเส้นทั่วไป (GLM) ดังที่กล่าวไปข้างต้น ในปี ค.ศ. 1972 Nelder และ Wedderburn ได้นำเสนอ ตัวแบบเชิงเส้นนัยทั่วไป หรือ Generalized linear model (GLMs) และถูกพัฒนาต่อมาโดย McCullagh และ Nelder ในปี ค.ศ. 1983 และ ค.ศ. 1989 หลังจากนั้นตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ได้เริ่มเป็นที่รู้จักและมีการนำมาประยุกต์ใช้ในสาขาต่างๆ อย่างแพร่หลายมากขึ้น ภายใต้การขยายผลและการพัฒนาใน 2 ประเด็นหลัก [2] ได้แก่ (1) การขยายผลของตัวแปรผลลัพธ์จากตัวแบบเชิงเส้นทั่วไป (GLM) เดิม ซึ่งถูกนำมาใช้ได้เฉพาะกับข้อมูลแบบต่อเนื่อง ให้มีความครอบคลุมข้อมูลแบบต่อเนื่องและแบบไม่ต่อเนื่อง หรือ ข้อมูลแบบปกติและแบบไม่ปกติ ภายใต้ฟังก์ชั่นการแจกแจงความน่าจะเป็น (p.d.f) ของตระกูลเอกโพเนนเชียล (Exponential family) เช่น Binomial, Poisson, Multinomial, Negative Binomial, Gamma และ Inverse gamma เป็นต้น และ (2) การพัฒนาแนวคิดการแปลงข้อมูลของกลุ่มตัวอย่างด้วย Log transformation ซึ่งมีข้อจำกัดดังกล่าวข้างต้น มาเป็นการแปลงข้อมูลในระดับประชากรจากฟังกชั่นการแจกแจงความน่าจะเป็น (p.d.f) ด้วยฟังก์ชั่นเชื่อมโยง (Link function) เพื่อเชื่อมโยงค่าเฉลี่ย

ของตัวแปรผลลัพธ์ให้อยู่ในรูปฟังก์ชั่นเชิงเส้นกับตัวแปรร่วมและการผันค่ากลับด้วยฟังก์ชั่นอินเวอร์ส (Inverse function) เพื่อหาค่าผลลัพธ์ หรือค่าประมาณของสัมประสิทธิ์ถดถอยที่ต้องการ ซึ่งสามารถสรุปเป็นประเด็น เพื่อจำแนกความแตกต่างระหว่างตัวแบบเชิงเส้นทั่วไป (GLM) และตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ได้ดังนี้

**ประเด็นที่ 1: ลักษณะข้อมูล หรือ การแจกแจงของตัวแปรผลลัพธ์** ตัวแบบเชิงเส้นทั่วไป (GLM) ใช้ได้กับเฉพาะตัวแปรผลลัพธ์แบบต่อเนื่องเท่านั้น ขณะที่ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) สามารถนำมาใช้ได้ครอบคลุมทั้งตัวแปรผลลัพธ์แบบต่อเนื่องและแบบไม่ต่อเนื่อง หรือ ตัวแปรผลลัพธ์ที่มีการแจกแจงแบบปกติและแบบไม่ปกติ

**ประเด็นที่ 2: วิธีการประมาณค่าพารามิเตอร์** ตัวแบบเชิงเส้นทั่วไป (GLM) ใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด หรือ Ordinary least square (OLS) ขณะที่ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด หรือ Maximum likelihood method (ML)

**ประเด็นที่ 3: การผ่อนคลายข้อตกลงเบื้องต้นที่เข้มงวด** สืบเนื่องจากตัวแบบทั้งสอง มีวิธีการประมาณค่าพารามิเตอร์ที่แตกต่างกัน จึงทำให้ความเข้มงวดของข้อตกลงเบื้องต้นของตัวแบบผ่อนคลายแตกต่างกันไปด้วย นั่นคือ ตัวแบบเชิงเส้นทั่วไป (GLM) ใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีการ OLS ซึ่งจำเป็นต้องอาศัยข้อตกลงเบื้องต้นที่ค่อนข้างเข้มงวดเกี่ยวกับการแจกแจงแบบปกติ, ความแปรปรวนคงที่และความเป็นอิสระ ขณะที่ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (ML) ซึ่งเป็นวิธีการที่พิจารณาค่าภาวะน่าจะเป็นบนพื้นฐานฟังก์ชั่นการแจกแจงความน่าจะเป็น (p.d.f) ของตัวแปรผลลัพธ์ที่ถูกนำมาพิจารณาโดยตรง ดังนั้นจึงทำให้นักวิจัยสามารถผ่อนคลาย หรือ ไม่จำเป็นต้องตรวจสอบข้อตกลงเบื้องต้นของความคลาดเคลื่อนอย่างเข้มงวด โดยเฉพาะการแจกแจงแบบปกติและความแปรปรวนท่ากัน หรือ คงที่ ดังกล่าว

ดังนั้นในทางปฏิบัติ ตัวแบบเชิงเส้นนัยทั่วไป (Generalized linear model, GLMs) จึงถือเป็นตัวแบบที่มีกรอบการทำงานหลักที่สำคัญประกอบด้วย ลักษณะข้อมูลตัวแปรผลลัพธ์ที่นำมาใช้ (ครอบคลุมทั้งตัวแปรผลลัพธ์แบบต่อเนื่องและแบบไม่ต่อเนื่อง หรือ ตัวแปรผลลัพธ์ที่มีการแจกแจงแบบปกติและไม่ปกติ) โดยมีส่วนประกอบ 3 ส่วนหลักที่ถูกนำมาพิจารณาเพื่อสร้างตัวแบบ (ได้แก่ ส่วนประกอบเชิงสุ่ม, ส่วนประกอบเชิงระบบและฟังก์ชั่นเชื่อมโยง) และมีข้อตกลงเบื้องต้นสำคัญที่ควรพิจารณา (ได้แก่ ขนาดตัวอย่างที่มากพอ, ค่าข้อมูลที่สูง หรือ ต่ำผิดปกติ และความเป็นอิสระต่อกัน) วิธีการประมาณค่าพารามิเตอร์ (ใช้วิธีภาวะน่าจะเป็นสูงสุด หรือ Maximum likelihood method, ML) และมีรูปสมการของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ดังนี้

❶.ส่วนประกอบเชิงสุ่ม → ❷.ฟังก์ชั่นเชื่อมโยง = ❸.ส่วนประกอบเชิงระบบ

$$Y \rightarrow g(.) = \eta \rightarrow \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

**❶. ส่วนประกอบเชิงสุ่ม (Random component) แทนด้วย** $Y$ หมายถึง เป็นส่วนที่พิจารณาและบ่งชี้ลักษณะการแจกแจงความน่าจะเป็นของความคลาดเคลื่อนจากข้อมูลตัวแปร

ผลลัพธ์ ภายใต้ฟังก์ชั่นการแจกแจงความน่าจะเป็นของตระกูลเอกซโพเนนเชียล โดยมีหลักการพิจารณาว่า “ หากรูปฟังก์ชั่นการแจกแจงความน่าจะเป็น (p.d.f.) สำหรับตัวแปรสุ่มแบบต่อเนื่องใด หรือ ฟังก์ชั่นมวลความน่าจะเป็น (p.m.f.) สำหรับตัวแปรสุ่มแบบไม่ต่อเนื่องใด สามารถแปลงรูปใหม่ให้สอดคล้องกับรูป p.d.f. ของ exponential family ได้ นั่นแสดงว่า รูปฟังก์ชั่น p.d.f. หรือ p.m.f. ดังกล่าวนั้น อยู่ในกลุ่มการแจกแจงในตระกูล Exponential family และสามารถนำคุณสมบัติที่มีอยู่ภายใต้การแจกแจงของตระกูลดังกล่าว ไปใช้ในการเทียบเคียงเพื่อสร้างและพัฒนาตัวแบบต่อไปได้ ”

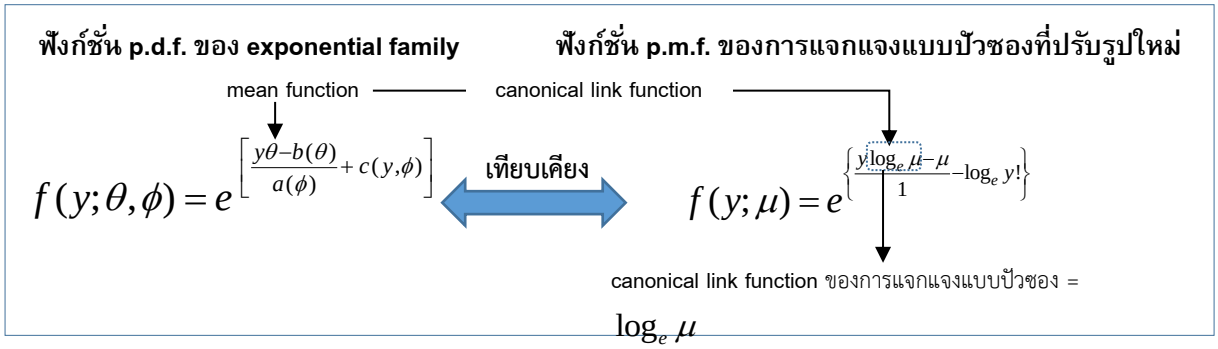
**ยกตัวอย่างเช่น** เมื่อฟังก์ชั่นการแจกแจงความน่าจะเป็นของความคลาดเคลื่อนภายใต้ตัวแปรผลลัพธ์ที่นักวิจัยสนใจศึกษามีการแจกแจงแบบปัวซอง (Poisson distribution) โดยมีรูปฟังก์ชั่น p.m.f เป็นดังนี้

$$f(y;\mu)=\frac{\mu^{y}e^{-\mu}}{y!}$$

จากรูปฟังก์ชั่น p.m.f ดังกล่าว นักวิจัยสามารถแปลงรูปฟังก์ชั่นใหม่ด้วยการ *take* $\log_e$ เข้าไปทั้งสองข้างของฟังก์ชั่น จะได้ $\log_e[f(y;\mu)]=y\log_e\mu-\mu-\log_e y!$ และปรับรูปของฟังก์ชั่น p.m.f ใหม่ เป็นดังนี้

$$f(y;\mu)=e^{\left\{\frac{y\log_e\mu-\mu}{1}-\log_e y!\right\}}$$

จากรูปฟังก์ชั่น p.m.f ใหม่ของการแจกแจงแบบปัวซอง พบว่า มีรูปแบบที่สามารถเทียบเคียงกับรูปฟังก์ชั่นการแจกแจงของตระกูล exponential family ได้ ดังนั้นจึงสามารถนำคุณสมบัติที่มีอยู่ของการแจกแจงตระกูล Exponential family มาใช้ประโยชน์ได้ โดยเฉพาะรูปของฟังก์ชั่นค่าเฉลี่ย (Mean function, $\theta$) ที่ได้จากการแจกแจงนี้ ซึ่งถูกนำไปใช้พิจารณาเป็นฟังก์ชั่นเชื่อมโยง (Link function) และเรียกว่า Canonical link function ดังนี้




**แผนภาพที่ 2** หลักการเทียบเคียงฟังก์ชั่นการแจกแจงภายใต้ตระกูล Exponential family เพื่อพิจารณา link function

ซึ่งในทางปฏิบัติ ส่วนประกอบเชิงสุ่มก็คือ รูปแบบการแจกแจงของข้อมูลตัวแปรผลลัพธ์ที่นำมาพิจารณานั่นเอง ซึ่งเมื่อทำการวิเคราะห์ด้วยโปรแกรมสำเร็จรูปทางสถิติ นักวิจัยจำเป็นต้องระบุให้สอดคล้องกับลักษณะข้อมูล ภายใต้รูปแบบคำสั่งของโปรแกรม ยกตัวอย่างเช่น ในโปรแกรม STATA การวิเคราะห์ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) จะใช้คำสั่ง **glm** โดยมีคำสั่งให้นักวิจัย

สามารถระบุส่วนประกอบเชิงสุ่ม ภายใต้ตัวเลือก **Family** เช่น กรณีข้อมูลต่อเนื่องและมีการแจกแจงแบบปกติ จะระบุเป็น **Family(gaussian)** หรือ กรณีข้อมูลแจงนับและมีการแจกแจงแบบทวินาม จะระบุเป็น **Family(binomial)** เป็นต้น

❷. **ส่วนประกอบเชิงระบบ (Systematic component) แทนด้วย $\eta$** ในทางปฏิบัติ ส่วนประกอบเชิงสุ่ม (Random component) จะถูกอธิบายความผันแปรเชิงสุ่มอย่างไม่เป็นระบบ (Unsystematic random variation) ภายใต้ค่าสังเกตที่นำมาวิเคราะห์ดังที่กล่าวไปข้างต้น ขณะที่ส่วนประกอบเชิงระบบ (Systematic component) จะถูกกำหนดให้เป็นส่วนโครงสร้างที่คงที่ (Fixed structural part) สำหรับตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ส่วนประกอบเชิงระบบ จะหมายถึง ตัวแปรทำนายเชิงเส้น (Linear predictor) หรือ บางครั้งเรียกว่า **“ผลรวมเชิงเส้นของตัวแปรทำนาย (Linear combination of predictors)”** แทนด้วยสัญลักษณ์ $\eta$ อ่านว่า “eta-อีต้า” โดยมีรูปสมการดังนี้

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +,...,+ \beta_k X_k$$

❸. **ฟังก์ชั่นเชื่อมโยง (Link function) แทนด้วย $g(.)$** เนื่องจากแนวคิดการทำงานของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ยังคงใช้ส่วนประกอบเชิงสุ่มในรูปตัวแปรทำนายเชิงเส้น (Linear combination) ดังกล่าวข้างต้น จึงทำให้ขอบเขตของค่าที่เป็นไปได้มีลักษณะเช่นเดียวกับตัวแบบเชิงเส้นทั่วไป (GLM) นั่นคือ มีค่าที่เป็นไปได้อยู่ในช่วง $[-\infty, \infty]$ หรือ เป็นได้ทั้งค่าติดลบ, ศูนย์และค่าบวก (จึงอาจสังเกตได้ว่า ตัวแบบนี้ยังคงใช้คำว่า “ตัวแบบเชิงเส้น” เนื่องจากมีส่วนประกอบเชิงสุ่มเป็น Linear combination อยู่นั่นเอง) และจากฟังก์ชั่นการแจกแจงในตระกูล Exponential family พบว่า คุณสมบัติของค่า $\theta$ (Canonical parameter) เป็นรูปฟังก์ชั่นค่าเฉลี่ย ซึ่งสามารถนำมาแปลงค่าขอบเขตของข้อมูลให้อยู่ในรูป $[-\infty, \infty]$ เพื่อเชื่อมโยงกับส่วนประกอบเชิงสุ่ม หรือ Linear combination ได้ ดังนั้นจึงกำหนดให้ $g(.)$ เป็นฟังก์ชั่นเชื่อมโยง (Link function) ภายใต้ค่าพารามิเตอร์ Canonical ($\theta$) หรือ รูปฟังก์ชั่นค่าเฉลี่ยที่ขึ้นกับรูปแบบฟังก์ชั่นการแจกแจงความน่าจะเป็นที่นำมาพิจารณาแต่ละกรณี โดยมีรูปทั่วไป ดังนี้

$$g(.) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +,...,+ \beta_k X_k$$

ขณะเดียวกันคุณสมบัติหนึ่งที่โดดเด่นของพารามิเตอร์ Canonical ($\theta$) หรือ รูปฟังก์ชั่นค่าเฉลี่ย ภายใต้ฟังก์ชั่นการแจกแจงในตระกูล Exponential family ซึ่งถูกนำมาใช้ในการกำหนดรูปฟังก์ชั่นเชื่อมโยง นั่นคือ การมีลักษณะฟังก์ชั่นแบบ Monotonic หรือ ฟังก์ชั่นแบบหนึ่งต่อหนึ่ง ดังนั้นจึงทำให้การแปลงค่าฟังก์ชั่นค่าเฉลี่ยไปในรูปสมการเชิงเส้นและการผันค่ากลับด้วยฟังก์ชั่น Inverse เพื่อหาค่าประมาณของค่าเฉลี่ย ซึ่งเป็นผลลัพธ์ในการนำไปใช้สรุปผลเพื่อตอบคำถามวิจัยสามารถทำได้อย่างตรงไปตรงมาและสะดวกในการนำมาใช้งานมากยิ่งขึ้น (ดังตารางที่ 1)

**ตารางที่ 1** แสดงส่วนประกอบเชิงสุ่ม ฟังก์ชั่นเชื่อมโยง ชื่อเรียกฟังก์ชั่นเชื่อมโยงและการผันค่ากลับด้วยฟังก์ชั่น Inverse ภายใต้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs)

| ส่วนประกอบเชิงสุ่ม | ฟังก์ชั่นเชื่อมโยง *** | ชื่อเรียกฟังก์ชั่นเชื่อมโยง | การผันค่ากลับด้วยฟังก์ชั่น inverse |
|---|---|---|---|
| Gaussian | $g(\mu)=\eta$ | Identity | $g^{-1}(\eta)=\mu$ |
| Binomial | $g(\log_e\left[\frac{\mu}{1-\mu}\right])=\eta$ | Logit | $g^{-1}(\eta)=\frac{e^{\alpha+\beta_1X_1+...+\beta_kX_k}}{1+e^{\alpha+\beta_1X_1+...+\beta_kX_k}}$ |
| Poisson | $g(\log_e\mu)=\eta$ | Log | $g^{-1}(\eta)=e^{\alpha+\beta_1X_1+...+\beta_kX_k}$ |

*** $\eta=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+,...,+\beta_kX_k$

**ความครอบคลุมและยืดหยุ่นของตัวแบบเชิงเส้นนัยทั่วไป (GLMs)**

จากกรอบการทำงานแบบหนึ่งเดียวของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ดังกล่าวข้างต้นได้สะท้อนค่อนข้างชัดเจนว่า วิธีการทางสถิติเกือบทั้งหมดที่นักวิจัยส่วนใหญ่คุ้นเคย ทั้งแบบต่อเนื่องซึ่งมีการแจกแจงแบบปกติและไม่ปกติ เช่น การวิเคราะห์ถดถอยพหุคูณ การวิเคราะห์แบบแกมมา เป็นต้น หรือ แบบไม่ต่อเนื่องทั้งแบบแจงนับและจำนวนนับ เช่น การวิเคราะห์ถดถอยลอจิสติก การวิเคราะห์ถดถอยปัวซอง เป็นต้น ล้วนถือเป็นวิธีการทางสถิติที่อยู่ภายใต้กรอบการทำงานของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ทั้งสิ้น เนื่องจากส่วนประกอบหลัก 3 ส่วนข้างต้น สามารถครอบคลุมแนวคิดพื้นฐานของวิธีการทางสถิติที่กล่าวมาทั้งหมด เช่น การวิเคราะห์ถดถอยพหุคูณจะให้ค่าผลลัพธ์เช่นเดียวกับการวิเคราะห์ข้อมูลด้วยตัวแบบเชิงเส้นนัยทั่วไป (GLMs) เมื่อกำหนดส่วนประกอบเชิงสุ่ม หรือ family ในคำสั่งโปรแกรม STATA เป็น Gaussian และกำหนดฟังก์ชั่นเชื่อมโยง หรือ Link function เป็น identity ขณะที่เช่นเดียวกันในการวิเคราะห์ถดถอยลอจิสติก ก็จะให้ค่าผลลัพธ์เดียวกันกับการวิเคราะห์ข้อมูลด้วยตัวแบบเชิงเส้นนัยทั่วไป (GLMs) เมื่อกำหนดส่วนประกอบเชิงสุ่ม หรือ Family ในคำสั่งโปรแกรม STATA เป็น Binomial และกำหนดฟังก์ชั่นเชื่อมโยง หรือ Link function เป็น Logit เป็นต้น จากตัวอย่างดังกล่าว นั่นแสดงว่า เมื่อนำตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มาวิเคราะห์ข้อมูล ภายใต้กรอบการทำงานแบบหนึ่งเดียวซึ่งครบทั้ง 3 ส่วนประกอบ จึงทำให้นักวิจัยสามารถวิเคราะห์ข้อมูลได้ครอบคลุมลักษณะข้อมูลของตัวแปรผลลัพธ์ที่มีอยู่และจากการที่มีส่วนประกอบฟังก์ชั่นเชื่อมโยง ยังทำให้นักวิจัยมีความยืดหยุ่นและสามารถเลือกรูปแบบฟังก์ชั่นเชื่อมโยงให้สอดคล้องกับลักษณะข้อมูลได้มากกว่าวิธีการทางสถิติโดยตรงแบบเดิม ซึ่งมีข้อจำกัดและทำไม่ได้ เช่น กรณีข้อมูลแบบต่อเนื่องและมีการแจกแจงแบบไม่ปกติ หากนำวิธีการทางสถิติเดิมภายใต้ตัวแบบเชิงเส้นทั่วไป (GLM) มาใช้ ได้แก่ สมการถดถอยเชิงพหุคูณ นักวิจัยจำเป็นต้องทำการแปลงค่าข้อมูล เพื่อให้สอดคล้องกับข้อตกลงเบื้องต้นที่มีอยู่และอาจนำไปสู่ข้อจำกัดจากการแปลงข้อมูลดังที่กล่าวไปข้างต้นได้ ขณะที่หากเลือกใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) นักวิจัยไม่จำเป็นต้องทำการแปลงข้อมูลตัวแปรผลลัพธ์ แต่สามารถเลือกตัวเลือกที่เป็นส่วนประกอบเชิงสุ่มและฟังก์ชั่นเชื่อมโยง ที่มีอยู่ในคำสั่งของโปรแกรมสำเร็จทางสถิติที่เลือกใช้ให้สอดคล้องและเหมาะสมกับข้อมูลจริงที่มีอยู่ เช่น ในกรณีโปรแกรม STATA จะพบว่า ทางเลือกของส่วนประกอบเชิงสุ่มกรณีข้อมูลต่อเนื่องทั้งแบบแจกแจงปกติและไม่ปกติ ประกอบด้วย Gaussian, Gamma, Inverse Gaussian และทางเลือกของฟังก์ชั่นเชื่อมโยง ประกอบด้วย Log, Power-1, Power-2 เป็นต้น

นอกจากความครอบคลุมและความยืดหยุ่นในประเด็นการแจกแจงของตัวแปรผลลัพธ์ภายใต้ส่วนประกอบเชิงสุ่มและการแปลงข้อมูลด้วยฟังก์ชั่นเชื่อมโยงแล้ว ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ยังได้มีการพัฒนาและขยายกรอบการทำงานแบบหนึ่งเดียวดังกล่าวอย่างต่อเนื่อง เพื่อให้เกิดความครอบคลุมสำหรับการวิเคราะห์ข้อมูลภายใต้คำถามวิจัยและลักษณะข้อมูลที่ความซับซ้อนและหลากหลายมากขึ้น [2] เช่น

**ตัวแบบ Generalized additive model (GAM)** ซึ่งพัฒนาและขยายต่อจากตัวแบบเชิงเส้นนัยทั่วไป (GLMs) เพื่อนำมาใช้ในกรณีที่นักวิจัยไม่แน่ใจเกี่ยวกับข้อตกลงเบื้องต้นประเด็นความสัมพันธ์เชิงเส้นตรง ดังนั้นวิธีการนี้ จึงเป็นทางเลือกในการนำเทคนิคเส้นโค้งราบเรียบ (Smoothing) ประกอบด้วยฟังก์ชั่น Spline หรือ Kernel smoothers มาใช้ในการสร้างกราฟเพื่ออธิบายการเปลี่ยนแปลงของข้อมูล โดยในทางปฏิบัติ นักวิจัยสามารถตัดสินใจเลือกใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) หรือ ตัวแบบ Generalized additive model (GAM) ขึ้นกับวัตถุประสงค์ หรือคำถามวิจัยเป็นหลัก โดยตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มีวัตถุประสงค์เพื่อมุ่งเน้นการประมาณค่า, การอ้างอิงและการแปลความหมายของค่าสัมประสิทธิ์ถดถอยที่ได้เป็นหลัก ขณะที่ตัวแบบ Generalized additive model ( GAM) มีวัตถุประสงค์ส่วนใหญ่เพื่อนำมาใช้ในการอธิบายแผนภาพ หรือ กราฟแนวโน้มเชิงเส้นโค้งที่เกิดขึ้นจากความสัมพันธ์ระหว่างตัวแปรผลลัพธ์กับตัวแปรร่วมที่นำมาพิจารณาเป็นหลัก [9]

**สมการประมาณค่านัยทั่วไป (Generalized estimating equation, GEE)** ซึ่งพัฒนาและขยายต่อจากตัวแบบเชิงเส้นนัยทั่วไป (GLMs) โดยในทางปฏิบัติวิธีการ GEE ไม่ถือเป็นตัวแบบทางสถิติ [10] เนื่องจากมีกรอบการทำงานเช่นเดียวกับตัวแบบเชิงเส้นนัยทั่วไป (GLMs) แต่พัฒนาวิธีการประมาณค่าพารามิเตอร์เพิ่มเติม เนื่องจากถูกนำมาใช้ในกรณีตัวแปรผลลัพธ์ที่สัมพันธ์กัน โดยคำนึงถึงและนำความสัมพันธ์ที่เกิดขึ้นภายในตัวแปรผลลัพธ์ที่สัมพันธ์กันดังกล่าวมาใช้ในการถ่วงน้ำหนักของการคำนวณค่าประมาณ เพื่อลดความลำเอียง (Bias) ที่อาจจะเกิดขึ้นจากความสัมพันธ์ภายในตัวแปรผลลัพธ์ดังกล่าว ด้วยการกำหนดโครงสร้างความสัมพันธ์แบบเมตริกซ์ หรือ Working correlation structure เพิ่มเติมเข้ามาในการคำนวณ ซึ่งหากการวิเคราะห์ข้อมูลด้วยวิธีการ GEE เมื่อนักวิจัยกำหนดโครงสร้างความสัมพันธ์ดังกล่าวได้อย่างถูกต้องและสอดคล้องกับโครงสร้างความสัมพันธ์ที่มีอยู่จริง (Correct specification) ผลของการประมาณค่าที่ได้ จะให้ค่าความคลาดเคลื่อนมาตรฐาน (Standard error) ที่มีความถูกต้องและแม่นยำ ส่งผลให้ทั้งค่าประมาณแบบจุด ผลการทดสอบสมมติฐานและช่วงเชื่อมั่นที่ได้ มีความแม่นยำและกระชับตามมา อย่างไรก็ตามแม้หากการกำหนดโครงสร้างความสัมพันธ์ดังกล่าวไม่ถูกต้อง หรือ ไม่สอดคล้องกับโครงสร้างความสัมพันธ์ที่มีอยู่จริง (misspecification) ผลของการประมาณค่าที่ได้ จะยังคงให้ค่าประมาณแบบจุดที่ไม่ลำเอียง (Unbias) อยู่ แต่จะส่งผลกระทบต่อความลำเอียงของค่าความคลาดเคลื่อนมาตรฐานและส่งผลต่อเนื่องทำให้การทดสอบสมมติฐานและช่วงเชื่อมั่นมีความไม่ถูกต้องและไม่แม่นยำตามมา ดังนั้นวิธีการแก้ปัญหาการเกิด Misspecification ดังกล่าว จึงแนะนำให้ใช้วิธีการประมาณค่าความคลาดเคลื่อนมาตรฐานแบบ Robust [9] ซึ่งนักวิจัยสามารถกำหนดเป็นทางเลือกได้ในโปรแกรมสำเร็จรูปทางสถิติที่นำมาใช้ทำการวิเคราะห์ข้อมูล

**ตัวแบบผลกระทบผสมเชิงเส้นนัยทั่วไป (Generalized linear mixed effects model, GLMM)** ถือเป็นตัวแบบหนึ่งที่มีความสำคัญ ซึ่งถูกพัฒนาและขยายต่อเนื่องมาจากตัวแบบเชิงเส้นนัยทั่วไป (GLMs) สำหรับการวิเคราะห์ข้อมูลที่ครอบคลุมลักษณะข้อมูลตัวแปรผลลัพธ์ได้ทั้งที่สัมพันธ์กันและเป็นอิสระต่อกัน ได้ทั้งแบบต่อเนื่องและไม่ต่อเนื่อง ภายใต้การนำผลกระทบทั้งแบบคงที่ (Fixed effect) และแบบสุ่ม (Random effect) มาพิจารณาร่วมกัน หรือเรียกว่า “ผลกระทบผสม (Mixed effects)” เพื่อตอบคำถามวิจัยได้ครอบคลุมทั้งแบบค่าเฉลี่ยประชากร (Population-averaged model) และแบบเฉพาะบุคคล (Subject-specific model)

บางครั้งนักวิจัยอาจสามารถสังเกตได้ว่าทุกครั้งที่พบเจอวิธีการทางสถิติ หรือ ตัวแบบทางสถิติ ที่ชื่อเรียกถูกนำหน้าด้วยคำว่า **“Generalize”** นั่นแสดงว่า วิธีการ หรือ ตัวแบบดังกล่าว ถูกพัฒนาและขยายผลมาจากกรอบการทำงานแบบหนึ่งเดียวของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) จึงทำให้สามารถนำมาใช้วิเคราะห์ข้อมูลได้ครอบคลุมทั้งตัวแปรผลลัพธ์แบบต่อเนื่องและแบบไม่ต่อเนื่อง หรือ มีการแจกแจงแบบปกติและแบบไม่ปกติ ดังนั้นจากประเด็นความครอบคลุมและยืดหยุ่นของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ที่กล่าวมาข้างต้น นักวิจัยจึงควรให้ความสำคัญและนำมาพิจารณา เพื่อประกอบการตัดสินใจเลือกใช้วิธีการวิเคราะห์ข้อมูลให้สอดคล้องและเหมาะสมกับคำถามวิจัยและลักษณะข้อมูลที่มีอยู่ โดยเฉพาะในงานวิจัยทางวิทยาศาสตร์สุขภาพ

## ข้อเสนอแนะการนำตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มาใช้ในงานวิจัยทางวิทยาศาสตร์สุขภาพ

1. ในงานวิจัยกรณีตัวแปรผลลัพธ์แบบต่อเนื่อง (Continuous outcome) เมื่อคำถามวิจัยต้องการทราบความสัมพันธ์ระหว่างตัวแปรผลลัพธ์กับตัวแปรร่วม ควรเลือกใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) จะมีความเหมาะสมมากกว่า เนื่องจากให้ความครอบคลุมและยืดหยุ่นในการจัดการตัวเลือกให้สอดคล้องกับลักษณะข้อมูลที่มีอยู่ได้ดีกว่าโดยสามารถเขียนระบุวิธีการวิเคราะห์ข้อมูลไว้ในโครงร่างการวิจัย (Proposal) ได้อย่างครอบคลุมมากกว่า เช่น “การวิเคราะห์ข้อมูลสำหรับการศึกษาครั้งนี้ ใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) โดยกรณีข้อมูลแบบต่อเนื่องที่มีการแจกแจงแบบปกติ จะกำหนดส่วนประกอบเชิงสุ่มเป็น Gaussian และฟังก์ชั่นเชื่อมโยงเป็น Identity ขณะที่หากข้อมูลแบบต่อเนื่องมีการแจกแจงแบบ Inverse Gaussian (เบ้ขวา) จะกำหนดส่วนประกอบเชิงสุ่มเป็น Igaussian และฟังก์ชั่นเชื่อมโยงเป็น Power -2” เป็นต้น

2. ในงานวิจัยกรณีตัวแปรผลลัพธ์แบบไม่ต่อเนื่อง (Discrete outcome) เช่นเดียวกัน นักวิจัยควรเลือกใช้ตัวแบบเชิงเส้นทั่วไป (GLMs) เพื่อให้เกิดความครอบคลุมและยืดหยุ่นในการวิเคราะห์ข้อมูลได้มากกว่า โดยเฉพาะในการเขียนโครงร่างการวิจัย (Proposal) ซึ่งส่วนใหญ่พบว่านักวิจัยยังไม่สามารถระบุได้อย่างแน่ชัดเกี่ยวกับข้อมูลที่จะถูกรวบรวมได้ว่า มีลักษณะสอดคล้องหรือ เป็นไปตามที่ระบุวิธีการวิเคราะห์ทางสถิติไว้หรือไม่ ? ดังนั้นการนำตัวแบบเชิงเส้นทั่วไป (GLMs) มาใช้ จึงสามารถเขียนแนวทางการวิเคราะห์ข้อมูลได้อย่างครอบคลุมและยืดหยุ่นมากยิ่งขึ้น เช่น การวิเคราะห์ข้อมูลตัวแปรผลลัพธ์แบบจำนวนนับ (Count outcome) ซึ่งถือเป็นข้อมูลที่มีการแจกแจงความน่าจะเป็นแบบปัวซองและวิธีการทางสถิติที่ใช้ ได้แก่ สมการถดถอยปัวซอง (Poisson regression) ภายใต้เงื่อนไขของ

การแจกแจงปัวซองที่ว่า ค่าเฉลี่ยและความแปรปรวนคงที่เท่ากัน ( $\mu = \sigma^2$ ) แต่ในทางปฏิบัติมักพบว่า ไม่เป็นไปตามเงื่อนไขดังกล่าว นั่นคือส่วนใหญ่ความแปรปรวนมากกว่าค่าเฉลี่ย หรือเรียกว่า “Overdispersion” ส่งผลให้การวิเคราะห์แบบเดิมด้วยสมการถดถอยปัวซอง จึงไม่เหมาะสมและจำเป็นต้องเลือกใช้สมการถดถอยทวินามลบ (Negative binomial regression) ซึ่งจากแนวทางดังกล่าว หากนักวิจัยเลือกตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มากำหนดวิธีการวิเคราะห์ ก็จะทำให้ง่ายและสะดวกในการพิจารณามากยิ่งขึ้น ดังเช่น “การวิเคราะห์ข้อมูลสำหรับการศึกษาครั้งนี้ ใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) โดยกรณีข้อมูลแบบจำนวนนับเป็นไปตามเงื่อนไขของการแจกแจงแบบปัวซอง ค่าเฉลี่ยและความแปรปรวนคงที่เท่ากัน ( $\mu = \sigma^2$ ) จะกำหนดส่วนประกอบเชิงสุ่มเป็น Poisson และฟังก์ชั่นเชื่อมโยงเป็น Log ขณะที่หากข้อมูลแบบจำนวนนับดังกล่าว มีความแปรปรวนมากกว่าค่าเฉลี่ย หรือ เกิด “Overdispersion” จะกำหนดส่วนประกอบเชิงสุ่มเป็น nbinomial และฟังก์ชั่นเชื่อมโยงเป็น log” เป็นต้น

3. ภายใต้กรอบการทำงานของตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ที่ครอบคลุมและยืดหยุ่น ยังทำให้นักวิจัยสามารถคำนวณหาค่าประมาณที่ต้องการได้มากขึ้น ด้วยการจับคู่ระหว่างส่วนประกอบเชิงสุ่ม (Family) กับฟังก์ชั่นเชื่อมโยง (Link function) ได้ เช่น ในการวิจัยทางวิทยาการระบาด นักวิจัยสามารถคำนวณหาค่า odds ratio ได้จากการกำหนด Family(binomial) และ Link(logit) และค่า Risk ratio ได้จากการกำหนด Family(binomial) และ Link(log) รวมถึงการหาค่า Risk different ได้จากการกำหนด Family(binomial) และ Link(identity) เป็นต้น [11]

4. การนำตัวแบบเชิงเส้นนัยทั่วไป (GLMs) มาใช้งาน ทำให้นักวิจัยมีความสะดวกและสามารถลดการเรียนรู้วิธีการทางสถิติแต่ละวิธีลงได้ และภายใต้งานวิจัยเดียวกัน ซึ่งอาจมีคำถามวิจัยย่อยแตกต่างกัน ก็สามารถเลือกส่วนประกอบเชิงสุ่มและฟังก์ชั่นเชื่อมโยงให้สอดคล้องไปตามบริบทของข้อมูลที่อยู่ภายใต้คำถามวิจัยย่อยที่ต้องการได้

**บทสรุป**

การวิเคราะห์ข้อมูลในงานวิจัยทางวิทยาศาสตร์สุขภาพ จำเป็นต้องอาศัยวิธีการทางสถิติที่มีความครอบคลุมและยืดหยุ่น เนื่องจากลักษณะข้อมูลและคำถามวิจัยที่มีอยู่ค่อนข้างจำเพาะและหลากหลายขึ้นกับบริบทในแต่ละสาขาที่เกี่ยวข้อง เช่น สาขาทางคลินิก ระบาดวิทยา หรืออนามัยสิ่งแวดล้อม เป็นต้น ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) จึงถือเป็นอีกหนึ่งทางเลือกที่มีกรอบการทำงานแบบหนึ่งเดียว ซึ่งครอบคลุมและยืดหยุ่น สามารถรองรับตัวแปรผลลัพธ์ได้ทั้งแบบต่อเนื่องและไม่ต่อเนื่อง รวมถึงยังสามารถพัฒนาตัวแบบที่ได้ให้สอดคล้องกับเงื่อนไข หรือคุณสมบัติของลักษณะข้อมูลที่เปลี่ยนแปลงไปได้ด้วยการเลือกฟังก์ชั่นเชื่อมโยงที่เหมาะสม นอกจากนี้ตัวแบบเชิงเส้นนัยทั่วไป (GLMs) ยังถูกพัฒนาและขยายไปเป็นตัวแบบทางสถิติ หรือวิธีการอื่น ซึ่งถูกนำไปใช้กับประเด็นอื่นได้อย่างครอบคลุมและต่อเนื่อง ดังนั้นการสนับสนุนให้นักวิจัยแต่ละสาขา หันมาสนใจและพิจารณานำตัวแบบดังกล่าวมาใช้ให้เพิ่มมากขึ้น จึงเป็นประเด็นจำเป็น เพื่อลดความซับซ้อนในการเรียนรู้และการใช้งานวิธีการทางสถิติแต่ละวิธี ซึ่งแตกต่างกันไปตามโปรแกรมสำเร็จรูป อีกทั้งยังเพิ่มโอกาสในการ

พัฒนาผลงานวิจัยให้เป็นที่ยอมรับและสามารถเผยแพร่ผลงานในระดับที่สูงขึ้นได้

**เอกสารอ้างอิง**


1. Casson, R.J. and L.D. Farmer, Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clinical & Experimental Ophthalmology, 2014. 42(6): p. 590-596.
2. Nelder, J.A., A large class of models derived from generalized linear models. Stat Med, 1998. 17(23): p. 2747-53.
3. Malehi, A.S., F. Pourmotahari, and K.A. Angali. Statistical models for the analysis of skewed healthcare cost data: a simulation study. Health economics review. 2015; 5: 11-11.
4. Feng, C., et al. Log-transformation and its implications for data analysis. Shanghai archives of psychiatry. 2014; 26(2): 105-109.
5. Keene, O.N. The log transformation is special. Statistics in Medicine. 1995; 14(8): 811-819.
6. Moran, J.L., et al. New models for old questions: generalized linear models for cost prediction. J Eval Clin Pract. 2007; 13(3): 381-9.
7. St-Pierre, A.P., V. Shikon, and D.C. Schneider. Count data in biology-Data transformation or model reformation? Ecology and evolution. 2018; 8(6): 3077-3085.
8. O'Hara, R.B. and D.J. Kotze. Do not log-transform count data. Methods in Ecology and Evolution. 2010; 1(2): 118-122.
9. Nitta, H., et al. An introduction to epidemiologic and statistical methods useful in environmental epidemiology. Journal of epidemiology. 2010; 20(3): 177-184.
10. Pekár, S. and M. Brabec. Generalized estimating equations: A pragmatic and flexible approach to the marginal GLM modelling of correlated data in the behavioural sciences. Ethology. 2018; 124(2): 86-93.
11. Naimi, A.I. and B.W. Whitcomb. Estimating Risk Ratios and Risk Differences Using Regression. American Journal of Epidemiology. 2020; 189(6): 508-510.