การหาฟังก์ชันเชื่อมโยงที่เหมาะสม สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood

A Study of Optimum Link Function for Quasi-likelihood Estimation

รุ้งรวี อำนาจตระกูล และ ลี่ลี อิงศรีสว่าง

Rungrawee Amnarttrakul and Lily Ingsrisawang

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ กรุงเทพฯ

บทคัดย่อ

การศึกษาครั้งนี้มีลักษณะเป็นแบบการทดลองโดยมีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood เมื่อข้อมูลมีลักษณะเป็นไปตามภายใต้หลักการของ GLM และเปรียบเทียบการใช้ฟังก์ชันเชื่อมโยงที่เหมาะสมในรูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ เมื่อไม่ทราบรูปแบบฟังก์ชันเชื่อมโยงภายใต้สถานการณ์ที่กำหนด สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood ในการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood จะกำหนดตัวแปรอิสระ 2 ตัวแปร คือ $X_1 \sim N(15,3^2)$ และ $X_2 \sim N(20,1.5^2)$ รูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ เมื่อ $\beta_0 = 10$, $\beta_1 = 5$ และ $\beta_2 = 2.5$ ซึ่ง Y จะมีการแจกแจงแบบปกติ ขนาดตัวอย่างที่ใช้ คือ 30 , 50 และ 100 ทำซ้ำจำนวน 1,000 ครั้ง พิจารณาค่า AIC ที่ได้จากวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุด และค่า QIC ที่ได้จากวิธีการประมาณค่าแบบ Quasi-likelihood พบว่า ค่าทั้งสองจะมีค่าเท่ากัน สรุปได้ว่า Likelihood function และ Quasi-likelihood function จะมีลักษณะเช่นเดียวกันในกรณีที่ข้อมูลมีการแจกแจงแบบปกติ สำหรับการเปรียบเทียบการใช้รูปแบบของฟังก์ชันเชื่อมโยงที่แตกต่างกัน รูปแบบฟังก์ชันเชื่อมโยงที่ใช้ในการศึกษา ได้แก่ รูปแบบเอกลักษณ์ รูปแบบกำลัง และและรูปแบบส่วนกลับ กำหนดตัวแปรอิสระ 4 ตัวแปร คือ $X_1 \sim N(15,3^2)$, $X_2 \sim N(20,1.5^2)$, $X_3 \sim B(1,0.3)$ และ $X_4 \sim B(1,0.7)$ รูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ เมื่อ $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = 2.5$, $\beta_3 = -1$ และ $\beta_4 = 3$ กำหนด ϕ มี 3 ระดับ ได้แก่ $\phi = 1$, 2 , 3 ขนาดตัวอย่างที่ใช้ คือ 30 , 50 และ 100 ทั้งหมด 27 สถานการณ์ ทำซ้ำจำนวน 1,000 ครั้ง แล้วทำการเก็บค่า Deviance เพื่อใช้เปรียบเทียบความเหมาะสมของรูปแบบความสัมพันธ์ที่สร้างขึ้น พบว่า ฟังก์ชันเชื่อมโยงในรูปแบบเอกลักษณ์จะให้ผลดีที่สุดในทุกขนาดตัวอย่าง ค่า Deviance ที่ได้จะมีค่าเพิ่มขึ้นตามค่า ϕ ที่เพิ่มขึ้นเช่นเดียวกันในทุกขนาดตัวอย่าง และเมื่อตัวอย่างขนาดใหญ่ขึ้นความแตกต่างของค่า Deviance จากฟังก์ชันเชื่อมโยงในแต่ละรูปแบบจะมีค่าแตกต่างกันน้อยลง

คำสำคัญ: การประมาณค่าแบบภาวะความน่าจะเป็นสูงสุด การประมาณค่าแบบ Quasi-likelihood ฟังก์ชันเชื่อมโยง

Abstract

The objectives of this study involved the comparison of two types of parameter estimations, maximum likelihood estimation (MLE) and Quasi-likelihood estimation (QLE), when the GLM was fitted to the data and involved the comparison of three types of link functions, identity link, power link, and reciprocal link, when the QLE was used and the link function was unknown. In order to compare the MLE with the QLE, 2 independent variables were defined, $X_1 \sim N(15,3^2)$ and $X_2 \sim N(20,1.5^2)$. The relationship between the dependent variable, Y, and the two independent variables was $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ when $\beta_0 = 10$, $\beta_1 = 5$, and $\beta_2 = 2.5$ when Y had a normal distribution. The sample sizes were set at 30, 50 and 100. For each sample size 1,000 simulation runs were made. The AIC from the MLE and the QIC from the QLE were compared to determine the difference between the MLE and the QLE. It was found that the likelihood function was equivalent to the Quasi-likelihood function when the data were from the normal distribution. For the comparison among three types of link functions, identity link, power link, and reciprocal link, 4 independent variables were defined, i-e, $X_1 \sim N(15,3^2)$, $X_2 \sim N(20,1.5^2)$, $X_3 \sim B(1,0.3)$ and $X_4 \sim B(1,0.7)$. The relationship between the dependent, Y, and the 4 independents was $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ when $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = 2.5$, $\beta_3 = -1$, and $\beta_4 = 3$. The levels of ϕ were 1, 2, and 3. The sample sizes were 30, 50 and 100. There were 27 situations and 1,000 simulation runs were made for each situation. The deviance used for judging the suitable link function was recorded from each run. It was found that when the QLE was used for all sample sizes, the most suitable link function was identity link and the deviance varied directly with ϕ. When the sample size was larger, the deviances from the link functions were slightly different.

Keywords: Maximum likelihood estimation , Quasi-likelihood estimation and Link function

1. บทนำ

เทคนิคการสร้างตัวแบบเชิงเส้นที่วางนัยทั่วไป (Generalized linear model) หรือที่เรียกว่า GLM เป็นเทคนิคทางสถิติเพื่อใช้สร้างรูปแบบความสัมพันธ์การระหว่างตัวแปรตาม (Dependent variable) กับตัวแปรอิสระ (Independent variables) เมื่อตัวแปรตามมีรูปแบบการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล (Exponential family) อันได้แก่ การแจกแจงแบบปกติ (Normal distribution) การแจกแจงแบบทวินาม (Binomial distribution) การแจกแจงแบบปัวซง (Poisson distribution) การ

และแจกแจงแบบเรขาคณิต (Geometric distribution) เป็นต้น ซึ่งสามารถใช้ได้กับรูปแบบความสัมพันธ์ที่เป็นทั้งแบบเชิงเส้น (Linear model) และแบบไม่เชิงเส้น (Nonlinear model) สำหรับการสร้างรูปแบบความสัมพันธ์ภายใต้หลักการของ GLM จะต้องคำนึงถึงสิ่งเหล่านี้ ประการแรก คือ การแจกแจงของข้อมูล (Response distribution) ประการที่สอง คือ รูปแบบความสัมพันธ์ระหว่างค่าเฉลี่ยของข้อมูล (Mean response) กับตัวแปรอิสระ [1]

สำหรับวิธีการประมาณค่าพารามิเตอร์ที่ใช้ภายใต้หลักการของ GLM ได้แก่ วิธีการภาวะความน่าจะเป็นสูงสุด (Maximum likelihood estimation) จะทำการสร้าง Likelihood function เพื่อนำไปหาตัวประมาณพารามิเตอร์แบบภาวะความน่าจะเป็นสูงสุด (Maximum likelihood estimator) [2] แต่ในทางปฏิบัติแล้วข้อมูลอาจจะไม่มีลักษณะการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล ซึ่งอาจจะมีลักษณะการแจกแจงเพียงคล้ายคลึงเท่านั้น หรือในกรณีที่เกิดปัญหาที่ข้อมูลมีการกระจายสูงกว่าปกติที่เรียกว่า Overdispersion [3] โดยจะพิจารณาได้จากพารามิเตอร์แสดงการกระจาย (Dispersion parameter : ϕ) หรือเรียกว่า Scale parameter เช่น กรณีที่ข้อมูลมีการแจกแจงแบบปกติ ϕ จะมีค่าเท่ากับ 1 แต่ถ้า ϕ มีค่ามากกว่า 1 แสดงว่าข้อมูลนี้เกิดปัญหา Overdispersion เมื่อนำวิธีการประมาณค่าแบบภาวะน่าจะเป็นสูงสุดมาประมาณค่าพารามิเตอร์จะทำให้ค่าคลาดเคลื่อนมาตรฐาน (Standard error) ของสัมประสิทธิ์การถดถอยมีค่าต่ำกว่าค่าที่ควรจะเป็น (Underestimated)

เมื่อข้อมูลมีการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียลจะสามารถเขียนฟังก์ชันความหนาแน่นน่าจะเป็น (Probability density function) ให้อยู่ในรูป

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

โดยที่ $a(.)$, $b(.)$ และ $c(.)$ เป็นฟังก์ชันของ ϕ , θ และ (y,ϕ) ตามลำดับ

θ เป็นพารามิเตอร์แสดงตำแหน่ง (Location parameter)

ϕ เป็นพารามิเตอร์แสดงการกระจาย (Dispersion parameter)

สำหรับ $a(.)$ จะอยู่ในรูป $a(\phi) = \phi.\omega$ เมื่อ ω เป็นค่าคงที่ ซึ่งการแจกแจงแบบทวินาม การแจกแจงแบบปัวซง และการแจกแจงแบบปกติ จะได้ว่า $\phi = 1$ [4]

จากรูปแบบฟังก์ชันความหนาแน่นน่าจะเป็นจะสามารถทำให้ทราบรูปแบบของฟังก์ชันตัวหนึ่ง ที่เรียกว่า ฟังก์ชันเชื่อมโยง (Link function) ซึ่งเป็นฟังก์ชันที่มีหน้าที่เชื่อมความสัมพันธ์ระหว่างค่าเฉลี่ยของข้อมูลกับตัวแปรพยากรณ์เชิงเส้น (Linear predictor) โดยเป็นองค์ประกอบหนึ่งที่สำคัญภายใต้หลักการ GLM เพื่อนำไปใช้สร้างรูปแบบความสัมพันธ์ของข้อมูล สำหรับในกรณีที่กำหนดให้ $\eta = g(\mu)$ เป็นฟังก์ชันเชื่อมโยงที่ทำให้ได้ตัวพยากรณ์เชิงเส้น หรือ $\eta = x'\underline{\beta}$ สามารถเป็นไปได้หลายรูปแบบ เช่น รูปแบบเอกลักษณ์ (Identity link) รูปแบบล็อก (Log Link) และรูปแบบกำลัง (Power link) เป็นต้น แต่โดยทั่วไปฟังก์ชันเชื่อมโยงที่มีรูปแบบ $\eta = \theta$ เรียกว่า Canonical link จะนิยมนำไปใช้ในการแปลงค่าเฉลี่ยของข้อมูล ซึ่งการเลือกใช้ Canonical link ไม่ได้หมายความว่าจะ

ให้ผลดีกว่าการเลือกฟังก์ชันเชื่อมโยงในรูปแบบอื่น [5] จากนั้นจึงทำการประมาณค่าพารามิเตอร์ของรูปแบบความสัมพันธ์ดังกล่าว โดยใช้วิธีการประมาณค่าแบบภาวะน่าจะเป็นสูงสุด ร่วมกับการคำนวณแบบวนซ้ำ (Iterative Technique)

ในกรณีที่ข้อมูลไม่มีลักษณะการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล หรือในกรณีที่เกิดปัญหา Overdispersion เช่น กรณีที่ข้อมูลมีการแจกแจงแบบปกติ การแจกแจงแบบทวินาม และการแจกแจงแบบปัวซง จะพบว่า $\phi > 1$ เมื่อนำวิธีการประมาณค่าแบบภาวะน่าจะเป็นสูงสุดมาประมาณค่าพารามิเตอร์จะทำให้ได้ตัวประมาณที่ไม่เหมาะสม ในปี ค.ศ.1974 Wedderburn [6] จึงได้เสนอวิธีการประมาณค่าพารามิเตอร์แบบ Quasi-likelihood ซึ่งเป็นวิธีการประมาณค่าพารามิเตอร์ที่เหมาะสมกับข้อมูลที่ไม่มีลักษณะการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล กรณีที่เกิดปัญหา Overdispersion หรือข้อมูลที่มีการวัดค่าซ้ำ (Repeated measure) ทั้งยังช่วยแก้ไขปัญหา Underestimated ซึ่งปัญหาสำคัญที่เกิดขึ้นอีกประการหนึ่งคือ ถ้าข้อมูลมีการแจกแจงที่ไม่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียลแล้วควรจะใช้ฟังก์ชันเชื่อมโยงในรูปแบบใดจึงจะเหมาะสม

งานวิจัยที่ศึกษาเกี่ยวกับวิธีการประมาณค่าแบบ Quasi-likelihood มีอยู่มากมาย ยกตัวอย่างเช่น Wedderburn [6] ได้เสนอว่าการสร้างรูปแบบความสัมพันธ์ทั้งแบบเชิงเส้นและไม่เชิงเส้น ในกรณีที่ความแปรปรวนของความคลาดเคลื่อนไม่คงที่ โดยใช้วิธีการประมาณค่าแบบ Quasi-likelihood ซึ่งจะใช้ Quasi-likelihood function แทนการใช้ Likelihood function ในวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุด และพบว่าถ้าทราบรูปแบบการแจกแจงของข้อมูล Log-likelihood function จะมีคุณสมบัติเช่นเดียวกับ Quasi-likelihood function สำหรับ McCullagh [7] ได้ศึกษาความเชื่อมโยงระหว่าง Quasi-likelihood function รูปแบบความสัมพันธ์ของข้อมูลที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล และรูปแบบความสัมพันธ์ไม่เชิงเส้นแบบถ่วงน้ำหนัก ซึ่งค่าประมาณที่ได้จะมีค่าเข้าใกล้ค่าพารามิเตอร์เมื่อตัวอย่างมีขนาดใหญ่ (Asymptotic) และ Firth [8] ได้ศึกษาโดยเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าแบบ Quasi-likelihood กับวิธีการประมาณค่าแบบภาวะน่าจะเป็นสูงสุด ซึ่งพบว่า วิธีการประมาณค่าแบบ Quasi-likelihood จะให้ประสิทธิภาพสูง เมื่อข้อมูลมีลักษณะการแจกแจงที่แตกต่างไปจากวงศ์ของเอ็กซ์โปเนนเชียล นอกจากนี้ Nelder และ Pregibon [9] ได้ขยายแนวคิดเกี่ยวกับ Quasi-likelihood function ต่อจากงานวิจัยของ Wedderburn โดยทำการเปรียบเทียบรูปแบบความสัมพันธ์ตามองค์ประกอบที่แตกต่างกันภายใต้หลักการของ GLM ได้แก่ ตัวพยากรณ์เชิงเส้น ฟังก์ชันเชื่อมโยง และฟังก์ชันของความแปรปรวน (Variance function) โดยถ้าข้อมูลมีการแจกแจงอยู่ในวงศ์ของเอ็กซ์โปเนนเชียลตามหลักการของ GLM จะสามารถใช้ Likelihood ratio และ Score tests มาทดสอบสมมติฐาน ซึ่งจะสามารถนำค่าเหล่านี้มาปรับใช้ให้เหมาะสมกับวิธีการประมาณค่าแบบ Quasi-likelihood ได้เช่นเดียวกัน Davidian และ Carroll [10] ทำการศึกษางานวิจัยของ Welder และ Pregibon ที่เกี่ยวกับวิธีการประมาณค่าแบบ Extended quasi-likelihood เพื่อให้ได้วิธีการประมาณค่าที่ทำให้ได้ตัวประมาณมีคุณสมบัติความคงเส้นคงวา

(Consistent) และไม่เอนเอียง (Unbiased) ซึ่งคุณสมบัติเหล่านี้เป็นคุณสมบัติเช่นเดียวกับในกรณีที่ข้อมูลมีการแจกแจงอยู่ในวงศ์เอ็กซ์โปเนนเชียล Hill และ Tsai [11] ได้ศึกษาวิธีการประมาณค่าแบบ Maximum quasi-likelihood โดยการกำหนดสถานการณ์ 2 กรณี ได้แก่ กรณีที่ทราบลักษณะการแจกแจงของข้อมูล ซึ่งจะทำการเปรียบเทียบระหว่างวิธีการแปลงข้อมูล (Transformation) กับ Quasi-likelihood ส่วนกรณีที่สอง คือ รูปแบบความสัมพันธ์ที่ทำการประมาณค่าแบบ Maximum likelihood แล้วมีความยุ่งยากและซับซ้อน ซึ่งพบว่า วิธีการประมาณค่าแบบ Maximum quasi-likelihood จะมีประสิทธิภาพดีกว่าวิธีการประมาณค่าแบบ Maximum Likelihood และการคำนวณสามารถทำได้ง่ายกว่า Weisberg และ Welsh [12] ได้พิจารณาการสร้างรูปแบบความสัมพันธ์ภายใต้หลักการ GLM เมื่อไม่ทราบรูปแบบของฟังก์ชันเชื่อมโยง โดยจะทำการประมาณค่าสัมประสิทธิ์การถดถอยโดยใช้ฟังก์ชันเชื่อมโยงในรูปแบบของ Canonical link แล้วจึงมาทำการประมาณฟังก์ชันเชื่อมโยงด้วยวิธีการทางนอนพาราเมตริก (Nonparametric) โดยใช้ตัวปรับให้เรียบของ Kernel ซึ่งตัวประมาณที่ได้จะมีคุณสมบัติความคงเส้นคงวา Chiou และ Muller [13] ได้ศึกษาวิธีการประมาณค่าแบบ Quasi-likelihood ในกรณีที่ไม่ทราบรูปแบบของฟังก์ชันเชื่อมโยงและฟังก์ชันของความแปรปรวน ซึ่งจะทำการประมาณ Quasi-likelihood function จากวิธีการทางนอนพาราเมตริกโดยการปรับให้เรียบ และนำฟังก์ชันที่ประมาณได้ไปใช้สำหรับการประมาณค่าพารามิเตอร์ ซึ่งวิธีการนี้จะทำให้การประมาณค่าพารามิเตอร์ทำได้ง่ายขึ้น [14] และในปี ค.ศ. 1999 ได้เสนอวิธีการประมาณค่าแบบ Nonparametric quasi-likelihood ซึ่งทำการประมาณความแปรปรวนจากค่ากำลังสองของความคลาดเคลื่อนโดยใช้วิธีทางนอนพาราเมตริก และได้เสนอว่าหลักเกณฑ์การเลือกค่า Bandwidth โดยพิจารณาค่า Deviance และ Pearson's Chi-Square [13]

ในงานวิจัยนี้ได้นำเสนอวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุด เมื่อข้อมูลเป็นไปตามภายใต้หลักการของ GLM เปรียบเทียบกับวิธีการประมาณค่าแบบ Quasi-likelihood เพื่อพิจารณาถึงความแตกต่างที่เกิดขึ้น และศึกษารูปแบบของฟังก์ชันเชื่อมโยงรูปแบบที่แตกต่างกัน ซึ่งในที่นี้จะพิจารณาในรูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ (Reciprocal link) โดยใช้วิธีการประมาณค่าแบบ Quasi-likelihood เพื่อพิจารณาถึงรูปแบบของฟังก์ชันเชื่อมโยงที่เหมาะสมกับข้อมูลที่ได้จำลองสถานการณ์ขึ้น

2. วิธีการทดลอง

การศึกษาครั้งนี้มีลักษณะเป็นแบบการทดลอง ทำการจำลองข้อมูลขึ้นด้วยการทำงานของเครื่องคอมพิวเตอร์โดยใช้โปรแกรม SAS แบ่งการทดลองออกเป็น 2 ส่วน ได้แก่

การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood เมื่อข้อมูลมีลักษณะเป็นไปตามภายใต้หลักการของ GLM ซึ่งมีรายละเอียดของวิธีการดังนี้

1. สร้างตัวแปรอิสระที่ใช้ในการทดลอง 2 ตัวแปร คือ $X_1 \sim N(15,3^2)$ และ $X_2 \sim N(20,1.5^2)$ เนื่องจากต้องการให้ตัวแปรตามมีการแจกแจงแบบปกติ ตัวแปรอิสระจึงต้องมีการแจกแจงแบบปกติ เพื่อให้สอดคล้องกับข้อสมมติภายใต้ของหลักการ GLM โดยสามารถเปลี่ยนจำนวนตัวแปร และค่าพารามิเตอร์ได้ ซึ่งในงานวิจัยนี้ได้แสดงผลการวิเคราะห์มาเพียง 1 กรณี กรณีอื่นๆสามารถทำได้ในทำนองเดียวกัน

2. สร้างตัวแปรตามที่ใช้ในการศึกษาภายใต้รูปแบบความสัมพันธ์เชิงเส้นเมื่อกำหนดค่าพารามิเตอร์ $\beta_0 = 10$, $\beta_1 = 5$ และ $\beta_2 = 2.5$ รูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ โดยค่าพารามิเตอร์สามารถเปลี่ยนเป็นค่าได้

3. กำหนดขนาดตัวอย่างเป็น 30 , 50 และ 100

4. สร้างรูปแบบความสัมพันธ์โดยใช้วิธีการประมาณค่าพารามิเตอร์แบบภาวะความน่าจะเป็นสูงสุดและวิธีการประมาณค่าแบบ Quasi-likelihood โดยใช้ฟังก์ชันเชื่อมโยงที่เป็น Canonical link รูปแบบเอกลักษณ์

5. ในแต่ละสถานการณ์ทำซ้ำจำนวน 1,000 ครั้ง

6. พิจารณาค่า AIC (Akaike's information criterion) ที่ได้จากวิธีการประมาณค่าพารามิเตอร์แบบภาวะความน่าจะเป็นสูงสุด

$$AIC = -2L + 2P$$

และค่า QIC (Quasi information criterion) ที่ได้จากวิธีการประมาณค่าแบบ Quasi-likelihood เมื่อขนาดตัวอย่างแตกต่างกัน

$$QIC = -2Q + 2P$$

โดยที่ L เป็น Log-likelihood function

Q เป็น Quasi-likelihood function

P เป็นจำนวนพารามิเตอร์ในตัวแบบ

เพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood

สำหรับการเปรียบเทียบการใช้รูปแบบของฟังก์ชันเชื่อมโยงที่แตกต่างกัน เมื่อไม่ทราบรูปแบบฟังก์ชันเชื่อมโยง สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood ซึ่งมีขอบเขตการศึกษาดังนี้

1. สร้างตัวแปรอิสระที่ใช้ในการทดลอง 4 ตัวแปร คือ $X_1 \sim N(15,3^2)$, $X_2 \sim N(20,1.5^2)$, $X_3 \sim B(1,0.3)$ และ $X_4 \sim B(1,0.7)$ เนื่องจากต้องการให้ตัวแปรอิสระประกอบด้วยตัวแปรชนิดต่อเนื่อง (Continuous) และไม่ต่อเนื่อง (Discrete) โดยสามารถเปลี่ยนจำนวนตัวแปร ลักษณะการแจกแจง และค่าพารามิเตอร์ได้ ซึ่งในงานวิจัยนี้ได้แสดงผลการวิเคราะห์มาเพียง 1 กรณี กรณีอื่นๆสามารถทำได้ในทำนองเดียวกัน

2. สร้างตัวแปรตามที่ใช้ในการศึกษาภายใต้รูปแบบความสัมพันธ์เชิงเส้น ในงานวิจัยนี้จะกำหนดค่าพารามิเตอร์ $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = 2.5$, $\beta_3 = -1$ และ $\beta_4 = 3$ รูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ โดยค่าพารามิเตอร์สามารถปรับเป็นค่าได้

3. กำหนด ϕ มี 3 ระดับ ได้แก่ $\phi = 1, 2, 3$

4. กำหนดขนาดตัวอย่างเป็น 30 , 50 และ 100

5. รูปแบบของฟังก์ชันเชื่อมโยงที่ใช้ในการศึกษา ได้แก่ รูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ

6. สร้างรูปแบบความสัมพันธ์โดยใช้วิธีการประมาณค่าแบบ Quasi-likelihood

7. ทำการวิเคราะห์ข้อมูลโดยวิธีการประมาณค่าพารามิเตอร์แบบ Quasi-likelihood ภายใต้ ϕ ขนาดตัวอย่าง และรูปแบบของฟังก์ชันเชื่อมโยง ทั้งหมด 27 สถานการณ์

8. ในแต่ละสถานการณ์ทำซ้ำจำนวน 1,000 ครั้ง

9. ทำการเก็บบันทึกค่า Deviance ที่ได้จากการสร้างรูปแบบความสัมพันธ์ตามสถานการณ์ที่กำหนด เพื่อหาฟังก์ชันเชื่อมโยงที่เหมาะสมสำหรับสถานการณ์ที่กำหนด โดยพิจารณาจากค่าเฉลี่ยของค่า Deviance

3. ผลการทดลองและวิจารณ์

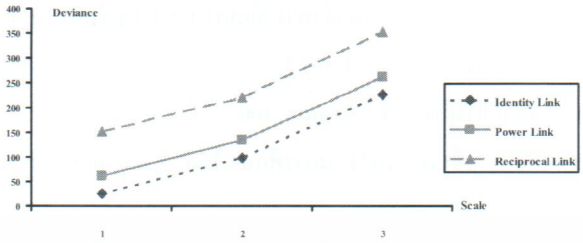
ผลการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-liklelihood เมื่อข้อมูลมีลักษณะเป็นไปตามภายใต้หลักการของ GLM พบว่า ให้ผลเช่นเดียวกัน ซึ่งแสดงว่า Quasi-likelihood function และ Likelihood function มีลักษณะเหมือนกัน ดังตารางที่ 3.1

ตารางที่ 3.1 ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของค่า AIC และ QIC

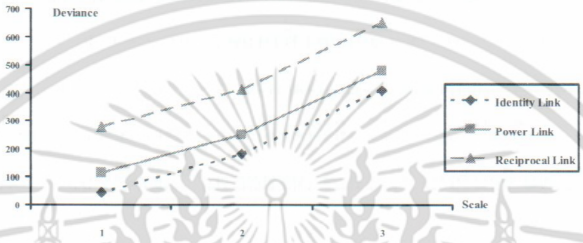
ขนาดตัวอย่าง	ค่า	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
30	AIC	4.5241	8.0946
	QIC	4.5241	8.0946
50	AIC	5.0518	10.195
	QIC	5.0518	10.1952
100	AIC	5.2223	14.4025
	QIC	5.2223	14.4025

ส่วนผลการเปรียบเทียบการใช้รูปแบบของฟังก์ชันเชื่อมโยง เมื่อไม่ทราบรูปแบบฟังก์ชันเชื่อมโยง สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood รูปแบบฟังก์ชันเชื่อมโยงที่ศึกษา ได้แก่

รูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ พิจารณาจากค่า Deviance ภายใต้สถานการณ์ที่กำหนด แสดงผลดังรูปที่ 3.1-3.3



รูปที่ 3.1 ค่าเฉลี่ยของ Deviance เมื่อขนาดตัวอย่าง 30 โดยกำหนด $\phi = 1, 2, 3$



รูปที่ 3.2 ค่าเฉลี่ยของ Deviance เมื่อขนาดตัวอย่าง 50 โดยกำหนด $\phi = 1, 2, 3$

รูปที่ 3.3 ค่าเฉลี่ยของ Deviance เมื่อขนาดตัวอย่าง 100 โดยกำหนด $\phi = 1, 2, 3$

จากรูปที่ 3.1-3.3 พบว่า ในทุกขนาดตัวอย่าง รูปแบบความสัมพันธ์ที่ใช้ฟังก์ชันเชื่อมโยงรูปแบบเอกลักษณ์จะให้ค่าเฉลี่ยของ Deviance ต่ำกว่าการใช้ฟังก์ชันเชื่อมโยงในรูปแบบอื่น และค่าเฉลี่ยของ Deviance ในทุกรูปแบบของฟังก์ชันเชื่อมโยงจะมีค่าเพิ่มขึ้นตามค่า ϕ

4. สรุปผลการทดลอง

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood สำหรับข้อมูลที่มีลักษณะเป็นไปตามภายใต้หลักการของ GLM และเปรียบเทียบการใช้ฟังก์ชันเชื่อมโยงที่เหมาะสมในรูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ เมื่อไม่ทราบรูปแบบฟังก์ชัน

เชื่อมโยงภายใต้สถานการณ์ที่กำหนด สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood โดยเป็นการศึกษาแบบการทดลอง ทำการจำลองข้อมูลด้วยโปรแกรม SAS ซึ่งผลของการทดลองจะแบ่งออกเป็น 2 ส่วน

ส่วนแรกเป็นการเปรียบเทียบวิธีการการประมาณค่าพารามิเตอร์ระหว่างวิธีการประมาณค่าแบบภาวะความน่าจะเป็นสูงสุดกับวิธีการประมาณค่าแบบ Quasi-likelihood โดยใช้ฟังก์ชันเชื่อมโยงที่เป็น Canonical link รูปแบบเอกลักษณ์ ซึ่งทำการสร้างตัวแปรอิสระ 2 ตัวแปร คือ $X_1 \sim N(15,3^2)$ และ $X_2 \sim N(20,1.5^2)$ สร้างตัวแปรตามที่ใช้ในการศึกษาภายใต้รูปแบบความสัมพันธ์เชิงเส้นโดยกำหนดค่าพารามิเตอร์ $\beta_0 = 10$, $\beta_1 = 5$ และ $\beta_2 = 2.5$ ในรูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ ซึ่ง Y จะมีการแจกแจงแบบปกติ เมื่อกำหนดขนาดตัวอย่างเป็น 30 , 50 และ 100 ในแต่ละสถานการณ์ทำซ้ำจำนวน 1,000 ครั้ง จากการพิจารณาค่า AIC ที่ได้จากวิธีการประมาณค่าพารามิเตอร์แบบภาวะความน่าจะเป็นสูงสุด และค่า QIC ที่ได้จากวิธีการประมาณค่าแบบ Quasi-likelihood พบว่า ในทุกขนาดตัวอย่างจะให้ค่าทั้งสองเท่ากัน แสดงว่า Quasi-likelihood function และ Likelihood function จะมีลักษณะเช่นเดียวกัน ในกรณีที่ข้อมูลมีการแจกแจงแบบปกติ หรือในกรณีที่ข้อมูลมีรูปแบบการแจกแจงที่อยู่ในวงศ์ของเอ็กซ์โปเนนเชียล

ในการเปรียบเทียบการใช้ฟังก์ชันเชื่อมโยงที่เหมาะสมในรูปแบบเอกลักษณ์ รูปแบบกำลัง และรูปแบบส่วนกลับ เมื่อไม่ทราบรูปแบบฟังก์ชันเชื่อมโยง สำหรับวิธีการประมาณค่าแบบ Quasi-likelihood ซึ่งทำการสร้างตัวแปรอิสระ 4 ตัวแปร คือ $X_1 \sim N(15,3^2)$, $X_2 \sim N(20,1.5^2)$, $X_3 \sim B(1,0.3)$ และ $X_4 \sim B(1,0.7)$ สร้างตัวแปรตามที่ใช้ในการศึกษาภายใต้รูปแบบความสัมพันธ์เชิงเส้นโดยกำหนดค่าพารามิเตอร์ $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = 2.5$, $\beta_3 = -1$ และ $\beta_4 = 3$ ในรูปแบบความสัมพันธ์ คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ กำหนด ϕ มี 3 ระดับ ได้แก่ $\phi = 1$, 2 , 3 ขนาดตัวอย่างเป็น 30 , 50 และ 100 ทำการวิเคราะห์ข้อมูลโดยวิธีการประมาณค่าพารามิเตอร์แบบ Quasi-likelihood ภายใต้ ϕ ขนาดตัวอย่าง และรูปแบบของฟังก์ชันที่แตกต่างกัน จำนวนทั้งหมด 27 สถานการณ์ ในแต่ละสถานการณ์ทำซ้ำจำนวน 1,000 ครั้ง การหาฟังก์ชันเชื่อมโยงที่เหมาะสมจะพิจารณาจากค่า Deviance ในรูปของค่าเฉลี่ย พบว่า กรณีขนาดตัวอย่าง 30 การใช้ฟังก์ชันเชื่อมโยงในรูปแบบเอกลักษณ์จะให้ผลดีที่สุด และค่า Deviance ที่ได้จะมีค่าเพิ่มขึ้นตามค่า ϕ ที่เพิ่มขึ้นเช่นเดียวกันในทุกขนาดตัวอย่าง และเมื่อตัวอย่างขนาดใหญ่ขึ้นความแตกต่างของค่า Deviance ที่จากการใช้ฟังก์ชันในแต่ละรูปแบบจะมีค่าแตกต่างกันน้อยลง

จากผลการศึกษาเป็นที่น่าสังเกตว่า ในกรณีที่ ϕ คงที่ และพิจารณาค่า Deviance เมื่อขนาดตัวอย่างมากขึ้น พบว่า ในแต่ละรูปแบบของฟังก์ชันเชื่อมโยงจะให้ผลที่แตกต่างกันมากขึ้น โดยอาจเป็นสิ่งที่บ่งบอกให้เห็นว่ารูปแบบของฟังก์ชันเชื่อมโยงที่แตกต่างกัน จะมีผลแตกต่างกันอย่างชัดเจนเมื่อตัวอย่างมีขนาดเพิ่มขึ้น ซึ่งเป็นเรื่องที่น่าสนใจศึกษาต่อไป

เอกสารอ้างอิง

[1] Myers, R.H., Montgomery, D.C. and Vining, G.G. 2002. Generalized Linear Models with Applications in Engineering and the Sciences., John Wiley, New York.

[2] Myers, R.H. and Milton, J.S. 1991. A First Course in the Theory of Linear Statistical Models., PWS-KENT, Boston.

[3] Gay, D.M. and Welsch, R.E. 1988. Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. J. Amer. Statist. Ass., 83, 990-998.

[4] McCullagh, P. and Nelder, J.A. 1996. Generalized Linear Models. Chapman&Hall, New York.

[5] Der, G. and Everitt, B.S. 2002. A Handbook of Statistical Analyses Using SAS. Chapman&Hall, New York.

[6] Wedderburn, R.W.M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61, 439-477.

[7] McCullagh, P. 1983. Quasi-likelihood functions. Ann. Statist., 11, 59-67.

[8] Firth, D. 1987. On the efficiency of quasi-likelihood estimation. Biometrika., 74, 233-245.

[9] Nelder, J.A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika, 74, 221-232.

[10] Davidian, M. and Carroll, R.J. 1988. A note on extended quasi-likelihood. Journal of the Royal Statistical Society. Series B (Methodological), 50, 74-82.

[11] Hill, J.R. and Tsai, C.L. 1988. Calculating the efficiency of maximum quasilikelihood estimation. App. Stat., 37, 219-230.

[12] Weisberg, S. and Welsh, A.H. 1994. Adaptive for the missing link. Ann Statist., 22, 1674-1700.

[13] Chiou, J.M. and Muller, H.G. 1999. Nonparametric quasi-likelihood. Ann. Statist., 27, 36-64.

[14] Chiou, J.M. and Muller, H.G. 1998. Quasi-likelihood regression with unknown link and variance functions. J. Amer. Statist. Ass., 93, 1376-1387.