ตัวแบบการแปลงค่าข้อมูลที่ให้ความแปรปรวนคงที่โดยประมาณ สำหรับข้อมูลปัวซง
Approximate Variance Stabilizing Transformation Model for Poisson Data

อุไรวรรณ เจริญกีรติกุล และ ลี่ลี อิงศรีสว่าง*
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
Uraiwan Jaroengeratikun and LiLy Ingsrisawang*
Department of Statistics, Faculty of Science, Kasetsart University.

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาตัวแบบการถดถอยสำหรับข้อมูลปัวซง โดยเปรียบเทียบระหว่างการสร้างตัวแบบถดถอยเชิงเส้นด้วยวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ และประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธีกำลังสองน้อยที่สุด (OLS) กับการสร้างตัวแบบถดถอยปัวซงจากข้อมูลโดยตรง และประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธีภาวะน่าจะเป็นสูงสุด (MLE) การศึกษานี้ทำการจำลองสถานการณ์ให้ตัวแปรตาม Y มีการแจกแจงแบบปัวซง และมีตัวแปรทำนาย 2 ตัว คือ X_1 และ X_2 กำหนดขนาดตัวอย่างที่ศึกษาเท่ากับ 10, 30, 50, 70 และ 100 โดย 1) สร้างฟังก์ชันการถดถอยเชิงเส้น $E(Y') = \mu_{Y'} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ ด้วยการแปลงค่า Y ใน 4 รูปแบบ คือ $Y' = \sqrt{Y}$, $Y' = \sqrt{Y+3/8}$, $Y' = \sqrt{Y+1} + \sqrt{Y}$ และ $Y' = \ln(Y+1)$ ตามลำดับ และ 2) สร้างฟังก์ชันการถดถอยปัวซง $E(Y) = \mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$ ทำให้มีสถานการณ์ที่แตกต่างกันทั้งหมด 25 สถานการณ์ ในแต่ละสถานการณ์จะถูกจำลองและประมาณค่าพารามิเตอร์ของตัวแบบด้วยโปรแกรม SAS® 9.1.3 โดยมีการคำนวณแบบวนซ้ำจำนวน 500 รอบ และพิจารณาความเหมาะสมของตัวแบบถดถอยที่สร้างขึ้นจากสถานการณ์ต่างๆ ด้วยค่าสถิติ Deviance ที่เฉลี่ยจากการวนซ้ำ 500 รอบ ($\overline{\text{Deviance}}$) ตัวแบบการถดถอยในสถานการณ์จำลองที่ให้ค่า $\overline{\text{Deviance}}$ ต่ำสุด จะเป็นตัวแบบการถดถอยที่เหมาะสมที่สุดสำหรับข้อมูลการแจกแจงปัวซง ผลการศึกษาพบว่าตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ (VST) ในรูปแบบ $Y' = \sqrt{Y+3/8}$ ให้ค่า $\overline{\text{Deviance}}$ ต่ำสุดเมื่อเทียบกับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงด้วยรูปแบบอื่น และให้ค่าใกล้เคียงกับค่า $\overline{\text{Deviance}}$ ของตัวแบบถดถอยปัวซง นอกจากนี้ผลการตรวจสอบความเหมาะสมของตัวแบบด้วยวิธีพล็อตค่าตกค้าง พบว่าค่าตกค้างจากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลในรูปแบบ $Y' = \sqrt{Y+3/8}$ มีลักษณะการกระจายสม่ำเสมอรอบๆ เส้น $Y' = 0$ ซึ่งเป็นไปตามคุณสมบัติของการประมาณค่าตัวแบบการถดถอยเชิงเส้น และยังพบว่าถ้าตัวอย่างข้อมูลมีขนาดใหญ่มากกว่าหรือเท่ากับ 50 ขึ้นไป ค่าทำนายของ Y ที่ได้จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลด้วย $Y' = \sqrt{Y+3/8}$ ให้ค่าที่ใกล้เคียงกับค่าทำนายจากตัวแบบถดถอยปัวซง แสดงว่าตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบ $Y' = \sqrt{Y+3/8}$ มีความเหมาะสมที่จะใช้เป็นตัวแบบสำหรับข้อมูลที่มีการแจกแจงแบบปัวซง

คำสำคัญ : การแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ ตัวแบบถดถอยปัวซง วิธีกำลังสองน้อยสุด (OLS) วิธีภาวะน่าจะเป็นสูงสุด (MLE)

Corresponding author. E-mail: fscilli@ku.ac.th

Abstract

The objective of this research was to study a regression model for Poisson data. Two types of regression models, including 1) a linear regression model that was applied for the variance stabilizing transformations and used the method of Ordinary Least Squares (OLS) for parameter estimates, and 2) a Poisson regression model in which its parameter estimates using the method of Maximum Likelihood Estimation (MLE) were considered and compared. The study method used a simulation technique. Data were simulated for the Poisson dependent variable, Y, and for the 2 predictor variables with the sample sizes of 10, 30, 50, 70, and 100 respectively. The simulation study consisted of : 1) building the linear regression model, $E(Y') = \mu_{Y'} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, in which Y was transformed in four patterns of $Y' = \sqrt{Y}$, $Y' = \sqrt{Y + 3/8}$, $Y' = \sqrt{Y+1} + \sqrt{Y}$, and $Y' = \ln(Y+1)$ respectively, and 2) building the Poisson regression model $E(Y) = \mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$. There were total 25 situations, and each situation 500 simulation runs were performed for parameter estimation by using SAS® 9.1.3. Additionally, the averaged value of deviance statistics that were obtained from the 500 simulation runs, denoted as $\overline{\text{Deviance}}$, was used for assessing the fit. The model with the smallest $\overline{\text{Deviance}}$ would be the most suitable model for Poisson data.

The results of this showed that the variance stabilizing transformation (VST) model in the form of $Y' = \sqrt{Y + 3/8}$ had the smallest $\overline{\text{Deviance}}$ among all types of the VST models and its value was still closed to the $\overline{\text{Deviance}}$ obtained from fitting the Poisson regression model. Moreover, the residual plot for model checking showed that the residuals fell within a horizontal band centered around 0 $(Y' = 0)$ with no systematic patterns. In addition, if the sample size was greater than or equal to 50, the predicted values of Y in the form of $Y' = \sqrt{Y + 3/8}$ was still closed to the ones obtained from the Poisson regression model. In conclusion, the approximate variance stabilizing transformation model in the form of $Y' = \sqrt{Y + 3/8}$ was suitable for Poisson data.

Keyword : Variance Stabilizing Transformation, Poisson Regression Model, Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE).

บทนำ

ปัจจุบันได้มีการพัฒนาตัวแบบถดถอยสำหรับข้อมูลจำนวนนับที่มีการแจกแจงแบบปัวซง (Poisson Distribution) มากขึ้น เช่น งานธุรกิจประกันภัยต้องการคาดหมายจำนวนครั้งการเรียกร้องค่าสินไหมทดแทน (Number of Claims) ของลูกค้าในกลุ่มอายุต่างๆ เพื่อนำผลการศึกษาที่ได้ไปใช้ในการจัดกลุ่มเสี่ยงภัย และกำหนดเบี้ยประกันภัยให้เกิดความยุติธรรมกับลูกค้า หรืองานบริการจราจรต้องการคาดหมายจำนวนรถยนต์ที่วิ่งบนถนนแต่ละเส้นทาง ณ ช่วงเวลาต่างๆ เพื่อกำหนดเวลาสับเปลี่ยนสัญญาณไฟจราจรที่เหมาะสม เป็นต้น การวิเคราะห์ตัวแบบสำหรับข้อมูลที่มีการแจกแจงแบบปัวซง $\mathrm{Y} = \mathrm{f}(\mathrm{x},\beta)+\varepsilon$; Y เป็นตัวแปรตาม (Dependent Variable) ที่มีการแจกแจงแบบปัวซง โดยมีค่าเฉลี่ย $\mathrm{f}(\mathrm{x},\beta) = \mu$ และความแปรปรวนของค่าคลาดเคลื่อน $\mathrm{Var}(\varepsilon)$ มีค่าไม่คงที่ (Heteroscedasticity) ซึ่งไม่เป็นไปตามคุณสมบัติของตัวแบบถดถอยเชิงเส้น (Linear Regression Model) ที่กำหนดว่า $\mathrm{Var}(\varepsilon)$ ต้องมีค่าคงที่ นอกจากนี้ความแปรปรวนของ Y ในตัวแบบถดถอยปัวซง ($\mathrm{Var}(\varepsilon)$) เป็นฟังก์ชันของค่าเฉลี่ย (Function of the Mean) คือ $\mathrm{Var}(\mathrm{Y}) = \mathrm{E}(\mathrm{Y}) = \mathrm{f}(\mathbf{x};\beta) = \mu = e^{\mathbf{x}^{T}\beta}$ เป็นค่าไม่คงที่ เมื่อ $\mathbf{x}$ มีค่าต่างๆ (McCullagh and Nelder, 1996; Myers *et al*., 2002; Kutner *et al*., 2005) ดังนั้นวิธีการจะทำให้เป็นไปตามคุณสมบัติตัวแบบถดถอยเชิงเส้น เพื่อจะได้ทำการประมาณค่าพารามิเตอร์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด (Ordinary Least Square: OLS) คือการใช้วิธีการแปลงข้อมูลด้วยรูปแบบ $\mathrm{Y}' = \sqrt{\mathrm{Y}}$ ทำให้ได้ $\mathrm{Var}(\mathrm{Y}')$ เป็นค่าคงที่ เรียกว่าเป็นการแปลงข้อมูลที่ทำให้ความแปรปรวนคงที่ (Variance Stabilizing Transformation: VST) (McCullagh and Nelder, 1996; Myers *et al*., 2002) ดังจะเห็นได้จาก Taylor Series ของ $\mathrm{Y}' = \sqrt{\mathrm{Y}}$ รอบเส้น $\mathrm{Y} = \mu$ โดยประมาณค่าด้วยสมการกำลังหนึ่ง (First Order Approximation): $\mathrm{Y}^{1/2} = \mu^{1/2} + \frac{d\mathrm{Y}^{1/2}}{d\mathrm{Y}}(\mathrm{Y}-\mu) = \mu^{1/2} + \frac{1}{2}\mu^{-1/2}(\mathrm{Y}-\mu)$ ได้ค่า $\mathrm{Var}(\mathrm{Y}^{1/2}) = \mathrm{Var}(\mu^{1/2} + \frac{1}{2}\mu^{-1/2}(\mathrm{Y}-\mu)) = \frac{1}{4}$ เป็นค่าคงที่ นอกจากนี้ Anscombe (1948) พบว่าการแปลงข้อมูลที่มีการแจกแจงแบบปัวซงด้วย $\mathrm{Y}' = \sqrt{\mathrm{Y}+3/8}$ ทำให้ $\mathrm{Var}(\mathrm{Y}')$ มีค่าคงที่เช่นเดียวกัน และหากข้อมูลปัวซง มีค่าเฉลี่ยมากกว่า 20 จะได้ $\mathrm{Var}(\mathrm{Y}')$ ลู่เข้าใกล้ค่า 1 ทำให้ข้อมูลที่ผ่านการแปลงค่าด้วย $\mathrm{Y}' = \sqrt{\mathrm{Y}+3/8}$ มีลักษณะการแจกแจงใกล้เคียงการแจกแจงปกติ (Normal Distribution) ส่วน Freeman and Tukey (1950) เสนอรูปแบบการแปลงข้อมูลที่ให้ $\mathrm{Var}(\mathrm{Y}')$ มีค่าคงที่ สำหรับกรณีที่มีข้อมูลบางค่าของ Y เท่ากับศูนย์ คือ $\mathrm{Y}' = \sqrt{\mathrm{Y}+1} + \sqrt{\mathrm{Y}}$ ต่อมา Bar-Lev and Enis (1988) ได้จัดกลุ่มรูปแบบการแปลงข้อมูลที่ทำให้ได้ VST โดยรวมวิธีการแปลงค่าข้อมูลของ Anscombe (1948) และ Freeman and Tukey (1950) พบว่าการแปลงข้อมูลปัวซงในรูปแบบ $\mathrm{Y}' = \sqrt{\mathrm{Y}+3/8}$ จะให้ค่า $\mathrm{Var}(\mathrm{Y}') = \frac{1}{4} + o(\mu^{-2})$ เมื่อ $\mu \to \infty$ หรือ $\mathrm{Var}(\mathrm{Y}') \approx \frac{1}{4}$ เป็นค่าคงที่ จนกระทั่งในปี ค.ศ. 2003 Rocke and Durbin ได้แสดงการแปลงข้อมูลจำนวนนับที่ทำให้ $\mathrm{Var}(\mathrm{Y}')$ มีค่าคงที่ ด้วยการแปลงในรูปแบบ Log-Linear ที่ข้อมูล Y ต้องมีค่าเฉลี่ย $\mathrm{E}(\mathrm{Y}) = \mu$ และความแปรปรวน $\mathrm{Var}(\mathrm{Y}) = a^2+b^2\mu^2$ ทำการแปลงข้อมูลในรูปแบบ $\mathrm{Y}' = \ln\left(\frac{\mathrm{Y}+\sqrt{\mathrm{Y}^2+c^2}}{2}\right)$ เมื่อ c=a/b, a และ b เป็นจำนวนจริง แต่ $b \neq 0$ เรียกรูปแบบนี้ว่า Generalized Logarithm (glog) เป็นรูปแบบการแปลงที่มีประโยชน์มากกว่าในรูปแบบการแปลง $\mathrm{Y} = \ln(\mathrm{Y}+c)$ เมื่อ c>0 แต่ถ้า Y มีค่าที่ใหญ่มาก (Extreme Value) จะได้ $\ln\left(\frac{\mathrm{Y}+\sqrt{\mathrm{Y}^2+c^2}}{2}\right) = \ln(\mathrm{Y})$ และ Uddin *et al*. (2006) ได้ให้ความสำคัญกับวิธีการแปลงค่าที่ทำให้ข้อมูลมีลักษณะการแจกแจงสมมาตร (Symmetry) หรือแจกแจงแบบลู่เข้าการแจกแจงปกติ เพื่อประโยชน์ในการอนุมานทางสถิติ โดยเฉพาะการประมาณค่าพารามิเตอร์ สำหรับข้อมูลจำนวนนับที่มีการแจกแจงแบบปัวซงการแปลงข้อมูลด้วยรูปแบบ $\mathrm{Y}' = \mathrm{Y}^{2/3}$ จะให้ลักษณะข้อมูลเป็นการแจกแจงแบบปกติมากกว่าการแปลงข้อมูลด้วยรูปแบบ $\mathrm{Y}' = \sqrt{\mathrm{Y}}$ ที่ทำให้ได้ $\mathrm{Var}(\mathrm{Y}')$ เป็นค่าคงที่ ต่อมา Lin *et al*. (2008) ได้เปรียบเทียบการแปลงข้อมูลจำนวนนับที่มีการวัดซ้ำด้วยรูปแบบ glog ที่ทำให้มีความแปรปรวนคงที่และมีการแจกแจงแบบปกติ (Variance Stabilizing Normalization: VSN) กับการแปลงข้อมูลด้วยรูปแบบ $\mathrm{Y}'' = a\mathrm{Y}' + b$ ที่เป็น VST เมื่อ Y มีค่าเฉลี่ย $\mathrm{E}(\mathrm{Y}) = \mu$ และความแปรปรวน $\mathrm{Var}(\mathrm{Y}) = v(\mu)$ โดยที่ $\mathrm{Var}(\mathrm{Y})$ เป็นฟังก์ชันของค่าเฉลี่ย (μ) และมี

$$Y' = \begin{cases} \dfrac{1}{c_1} \operatorname{arcsin h}\left(\dfrac{c_2}{\sqrt{c_3}} + \dfrac{c_1 Y}{\sqrt{c_3}}\right) & ; c_3 > 0 \\ (1/c_1)\ln(c_2 + c_1 Y) & ; c_3 = 0 \end{cases}$$

โดยที่ c_1 , c_2 และ c_3 หาได้จากสมการ $\sqrt{\nu(\mu) - c_3} = c_1\mu + c_2$; ค่า c_3 ประมาณจากค่าเฉลี่ยของความแปรปรวน Y ของแต่ละหน่วยสังเกต สำหรับค่า c_1 และ c_2 ประมาณจากวิธีการประมาณค่าตัวแบบถดถอยเชิงเส้นพบว่า ในกรณีข้อมูลจำนวนนับที่มีการวัดซ้ำการแปลงด้วย Y'' จะให้ค่า $\mathrm{Var}(Y'')$ คงที่ที่มีประสิทธิภาพดีกว่าการแปลงข้อมูลด้วยรูปแบบ glog

จากงานวิจัยที่เกี่ยวข้องนี้มีการแปลงข้อมูลที่เข้าข่ายลักษณะ VST หลากหลายวิธี การเลือกวิธีการแปลงข้อมูลที่มีลักษณะ VST ที่เหมาะสมจึงมีความสำคัญต่อการประมาณค่า Y ในงานวิจัยนี้จึงสนใจตัวแบบถดถอยที่เหมาะสมกับข้อมูลปัวซงโดยเปรียบเทียบระหว่างตัวแบบถดถอยเชิงเส้นที่ข้อมูลมีการแปลงด้วยรูปแบบต่างๆ ให้มีความแปรปรวนคงที่และประมาณค่าพารามิเตอร์ด้วย OLS กับตัวแบบถดถอยปัวซงที่ได้จากข้อมูลปัวซงโดยตรงและประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation : MLE)

วัตถุประสงค์

เพื่อศึกษาตัวแบบการถดถอยสำหรับข้อมูลปัวซง โดยเปรียบเทียบระหว่างการสร้างตัวแบบถดถอยเชิงเส้นด้วยวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ และประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธี OLS กับการสร้างตัวแบบถดถอยปัวซงจากข้อมูลโดยตรง และประมาณค่าพารามิเตอร์ด้วยวิธี MLE

ขอบเขตการวิจัย

การศึกษานี้ใช้การจำลองสถานการณ์ โดย Y เป็นตัวแปรตาม หรือ ตัวแปรตอบสนอง (Response Variable) แทนข้อมูลจำนวนนับที่มีการแจกแจงแบบปัวซงด้วยฟังก์ชันมวลความน่าจะเป็น (Probability Mass Function: PMF) คือ $\Pr(Y = y) = \dfrac{e^{-\mu}\mu^y}{y!}$; $y = 0,1,2,...$ โดยมีค่าเฉลี่ย $E(Y) = \mu$ และความแปรปรวน $\mathrm{Var}(Y) = \mu$

1 สร้างตัวแบบถดถอยปัวซงของข้อมูลจำนวนนับโดยตรงด้วยรูปแบบสมการ $E(Y) = \mu = f(\mathbf{x};\boldsymbol{\beta}) = e^{\mathbf{x}^T\boldsymbol{\beta}}$ ใช้ฟังก์ชันเชื่อมโยง (Link Function) แบบ $\ln(\mu) = \mathbf{x}^T\boldsymbol{\beta}$ เพื่อให้ตัวแบบถดถอยมีความสัมพันธ์แบบเชิงเส้น และทำการประมาณค่าพารามิเตอร์ด้วยวิธี MLE

2. สร้างตัวแบบถดถอยเชิงเส้น โดยการแปลงค่า Y ด้วย 4 รูปแบบ คือ $Y' = \sqrt{Y}$, $Y' = \sqrt{Y + 3/8}$, $Y' = \sqrt{Y+1} + \sqrt{Y}$ และ $Y' = \ln(Y+1)$ ที่ให้สมการ $E(Y') = \mu_{Y'} = \mathbf{x}^T\boldsymbol{\beta}$ และทำการประมาณค่าพารามิเตอร์ด้วยวิธี OLS

3. การจำลองสถานการณ์ตัวแบบถดถอยปัวซงในข้อ 3.1 และตัวแบบถดถอยเชิงเส้นในข้อ 2

3.1 กำหนดตัวแปร X_1 และ X_2 เป็นตัวแปรทำนาย (Predictor Variables) โดยที่ตัวแปร X_1 มีค่า 10 ระดับ เท่ากับ 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5 และ 5 และตัวแปร X_2 มีค่าต่อเนื่องในช่วง (1, 10) ค่าของตัวแปร X_1 และ X_2 ที่นำมาศึกษาเป็นระดับค่าที่ใช้ประยุกต์ในงานด้านต่างๆ เช่น ตัวอย่างงานด้านประกันภัยรถยนต์ในประเทศไทย ที่มีปัจจัยค่าส่วนลดเบี้ยประกันภัยและระยะทางการวิ่งต่อเที่ยว มีผลต่อจำนวนครั้งการเรียกร้องสินไหมทดแทน โดยลักษณะปัจจัยค่าส่วนลดเบี้ยประกันภัย (หน่วยเป็น 10%) จะปรับเพิ่มส่วนลดครั้งละ 0.5 ซึ่งให้ค่าต่ำสุดเท่ากับ 0.5 และค่าสูงสุดเท่ากับ 5 และลักษณะปัจจัยระยะทางการวิ่งต่อเที่ยว จะเป็นค่าต่อเนื่อง (หน่วย 10 กิโลเมตร)

ส่วนตัวแปรตาม Y กำหนดให้มีค่าเป็นจำนวนนับที่มีการแจกแจงแบบปัวซง

3.2 ขนาดตัวอย่างที่ศึกษา (Sample Sizes: n) เท่ากับ 10, 30, 50, 70 และ 100

3.3 การจำลองข้อมูลให้มีการแจกแจงปัวซงจะใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ด้วยโปรแกรม SAS® 9.1.3 ทำการจำลองสถานการณ์ทั้งหมด 25 สถานการณ์ และในแต่ละสถานการณ์จะมีคำนวณแบบวนซ้ำจำนวน 500 รอบ เพื่อประมาณค่าพารามิเตอร์ของตัวแบบ

วิธีการดำเนินการวิจัย

1 ทำการจำลองข้อมูลให้มีการแจกแจงแบบปัวซงของแต่ละสถานการณ์และนำข้อมูลมาศึกษา 2 ตัวแบบ

1.1 ตัวแบบ $Y = f(x,\beta)+\varepsilon$; Y มีการแจกแจงแบบปัวซง โดยมีค่าเฉลี่ย คือ $E(Y) = f(\mathbf{x};\beta) = \mu$ เท่ากับ $e^{\mathbf{x}^T\boldsymbol{\beta}} = e^{\beta_0+\beta_1X_1+\beta_2X_2}$ ที่มีค่าประมาณหรือค่าทำนาย (Predicted Value) ของ Y เท่ากับ

$$\hat{Y} = \hat{f}(\mathbf{x},\boldsymbol{\beta}) = \hat{\mu} = e^{\mathbf{x}^T\mathbf{b}} = e^{b_0+b_1X_1+b_2X_2} \quad (1)$$

เมื่อ $\mathbf{x}^T = [1 \quad X_1 \quad X_2]$ และ **b** เป็นเวกเตอร์ของตัวประมาณ

ค่าพารามิเตอร์ในตัวแบบ โดยหาค่า b_0, b_1 และ b_2 ด้วยวิธี MLE

พิจารณา log likelihood function ของ $Y_1, Y_2, Y_3,..., Y_n$ ที่ Y_i เป็นอิสระต่อกัน (Myer *et al.*, 2002) คือ

$$\ln L(\boldsymbol{\beta};\mathbf{y}) = \ln\prod_{i=1}^{n}\left(\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right) = \sum_{i=1}^{n}(e^{-\mathbf{x}_i^T\boldsymbol{\beta}} + y_i\mathbf{x}_i^T\boldsymbol{\beta} - \ln y_i!)$$

ทำการอนุพันธ์อันดับที่ 1 (First Derivative) เทียบกับ $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$ และกำหนดให้เท่ากับ 0

$$\frac{\partial \ln L(\mathbf{y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i - e^{\mathbf{x}_i^T\boldsymbol{\beta}})\cdot\mathbf{x}_i = 0 \qquad (2)$$

หาค่า b_0, b_1 และ b_2 โดยวิธีการคำนวณแบบวนซ้ำ (Numerical Iterative) ในสมการ (2) จนได้ค่าประมาณที่ลู่เข้าค่าๆ หนึ่ง ก็จะได้ค่า b_0, b_1 และ b_2 เป็นตัวประมาณค่าแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator) ของ β_0 , β_1 และ β_2 ตามลำดับ และค่าประมาณความคลาดเคลื่อนมาตรฐาน (Standard Error) ของ **b** เท่ากับ $SE(\mathbf{b}) = (\mathbf{x}^T\hat{V}\mathbf{x})^{-1/2}$ เมื่อ $\hat{V} = diag(e^{\mathbf{x}^T\mathbf{b}})$

1.2 จากตัวแบบในข้อ 1.1 ทำการแปลงข้อมูล Y ด้วย 4 รูปแบบ คือ $Y' = \sqrt{Y}$, $Y' = \sqrt{Y+3/8}$, $Y' = \sqrt{Y+1} + \sqrt{Y}$ และ $Y' = \ln(Y+1)$ ทำให้ได้ตัวแบบเชิงเส้น $Y' = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ $= \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$ ที่มีคุณสมบัติตามข้อตกลง คือ ค่าความแปรปรวนของ Y' และ ε มีค่าคงที่ และสามารถประมาณค่าพารามิเตอร์ด้วย OLS คือ $\mathbf{b} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^TY'$ และได้ค่าประมาณหรือค่าทำนายของ Y' เท่ากับ

$$\hat{Y}' = \mathbf{x}^T\mathbf{b} = b_0 + b_1X_1 + b_2X_2 \qquad (3)$$

ที่มี $SE(\mathbf{b}) = (\mathbf{x}^T\mathbf{x})^{-1}\cdot s^2$ โดยที่ s^2 เป็นค่าประมาณของ $Var(\varepsilon) = \sigma^2$ ในตัวแบบของ Y' ซึ่ง s^2 ก็คือค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (Mean Squared Error : MSE) มีค่าเท่ากับ $\sum_{i=1}^{n}\frac{e_i^2}{n-p}$ เมื่อ p=3 เป็นจำนวนพารามิเตอร์ในตัวแบบ และ $e_i = Y_i' - \hat{Y}_i'$ เป็นค่าตกค้าง (Residual) (Myers and Milton, 1991; Kutner *et al.*, 2005)

2 เกณฑ์การเปรียบเทียบความเหมาะสมระหว่างตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูล Y เพื่อให้ได้ความแปรปรวนคงที่ใน 4 รูปแบบ กับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y ด้วยค่า Deviance ที่เฉลี่ยได้จากการวนซ้ำ ($\overline{\text{Deviance}}$) คือ

$$\overline{\text{Deviance}} = \frac{\sum_{k=1}^{N} D(\boldsymbol{\beta})_k}{N} \qquad (4)$$

เมื่อ k คือรอบที่วนซ้ำ, N คือจำนวนรอบของการคำนวณแบบวนซ้ำในที่นี้กำหนด N=500 และ $D(\boldsymbol{\beta})_k$ คือค่า Deviance จากตัวแบบถดถอยในรอบการวนซ้ำที่ k (Myers *et al.*, 2002) คำนวณได้ดังนี้

$$\begin{aligned} D(\beta)_k &= -2\ln\frac{L(\boldsymbol{\beta})}{L(\mu)} \\ &= 2\left(\ln L(\mu) - \ln L(\beta)\right) \\ &= 2\left[-\sum_{i=1}^{n}(y_i - \hat{\mu}_i) + \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)\right] \end{aligned} \qquad (5)$$

เมื่อ ln L(β) เป็น log likelihood ของตัวแบบถดถอยปัวซงที่มีค่าเท่ากับ

$$-\sum_{i=1}^{n}\hat{\mu}_i + \sum_{i=1}^{n} y_i \ln(\hat{\mu}_i) - \sum_{i=1}^{n}\ln y_i!$$

ln L(μ) เป็น log likelihood ของ Saturated Model ที่ไม่มีตัวแปรทำนายใดๆ ในตัวแบบ ที่มีค่าเท่ากับ

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \ln y_i - \sum_{i=1}^{n}\ln y_i!$$

และค่า $\hat{\mu}$ เป็นค่าทำนายของ Y

สำหรับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y โดยตรงจะมีค่า $\hat{\mu}$ เท่ากับ $\hat{Y} = \hat{f}(\mathbf{x},\boldsymbol{\beta}) = e^{b_0+b_1X_1+b_2X_2}$ ตามสมการ (1)

และสำหรับตัวแบบถดถอยเชิงส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ที่มีค่าประมาณ Y' ตามสมการ (3) จะมีค่าทำนายของ Y ($\hat{\mu}$) ขึ้นกับวิธีการแปลงข้อมูลในรูปแบบต่างๆ ดังนี้

- วิธีการแปลงแบบ $Y' = \sqrt{Y}$ มีค่า $\hat{\mu} = (\hat{Y}')^2$

- วิธีการแปลงในรูปแบบ $Y' = \sqrt{Y+3/8}$ มีค่า $\hat{\mu} = (\hat{Y}')^2 - 0.375$

- วิธีการแปลงในรูปแบบ $Y' = \sqrt{Y+1} + \sqrt{Y}$ มีค่า $\hat{\mu} = \left(\frac{(\hat{Y}')^2 - 1}{2\hat{Y}'}\right)^2$

- วิธีการแปลงในรูปแบบ $Y' = \ln(Y+1)$ มีค่า $\hat{\mu} = e^{\hat{Y}'} - 1$

เกณฑ์การพิจารณาตัวแบบถดถอยที่มีความเหมาะสมกับข้อมูลคือ ตัวแบบที่ให้ค่า $\overline{\text{Deviance}}$ มีค่าต่ำสุด

3 การตรวจสอบตัวแบบถดถอยสำหรับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ ใช้วิธีพล็อตค่าตกค้าง (Residual Plot) กับค่าทำนายของ Y'' (Kutner *et al*., 2005) และในการตรวจสอบตัวแบบสำหรับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y โดยตรง ใช้วิธี Deviance Residuals Plot กับค่าทำนายของ Y $(\hat{\mu} = e^{x^T b})$ เมื่อค่า Deviance Residual เท่ากับ

$$\text{sign}(Y_i - \hat{\mu}_i)\sqrt{2Y_i(\ln Y_i - \ln \hat{\mu}_i)} \quad (6)$$

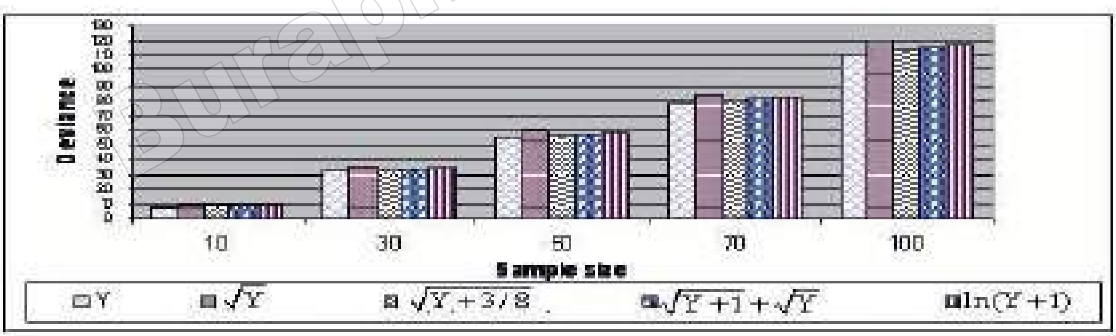
(Myers *et al*., 2002)

ผลการวิจัย

1 ผลศึกษาตัวแบบการถดถอยสำหรับข้อมูลปัวซง โดยเปรียบเทียบระหว่างการสร้างตัวแบบถดถอยเชิงเส้นด้วยวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ และประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธี OLS กับการสร้างตัวแบบถดถอยปัวซงจากข้อมูลโดยตรง และประมาณค่าพารามิเตอร์ด้วยวิธี MLE ที่ใช้เกณฑ์การเปรียบเทียบจากค่า Deviance ที่เฉลี่ยจากการวนซ้ำ ตามแสดงในตารางที่ 1 และภาพที่ 1 แสดงให้เห็นว่าทุกสถานการณ์ตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ การแปลงด้วยรูปแบบ $Y' = \sqrt{Y + 3/8}$ ให้ค่า $\overline{\text{Deviance}}$ ต่ำสุด เมื่อเทียบกับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงด้วยรูปแบบอื่น และให้ค่าใกล้เคียงกับค่า $\overline{\text{Deviance}}$ ของตัวแบบถดถอยปัวซง ดังนั้นถ้าจะต้องเลือกใช้วิธีการแปลงข้อมูลปัวซง ในที่นี้เสนอให้เลือกใช้ตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบ $Y' = \sqrt{Y + 3/8}$ เป็นตัวแบบสำหรับข้อมูลปัวซง

ตารางที่ 1 ค่า Deviance ที่เฉลี่ยจากการวนซ้ำ 500 รอบของตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y และตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลลักษณะ VST ในรูปแบบต่างๆ

ขนาดตัวอย่าง (n)	ตัวแบบถดถอยปัวซงจาก Y	ตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลลักษณะ VST ในรูปแบบ			
		$\sqrt{Y}$	$\sqrt{Y+3/8}$	$\sqrt{Y+1}+\sqrt{Y}$	$\ln(Y+1)$
10	8.149	9.101	8.340	8.597	8.937
30	31.736	34.758	32.590	33.317	33.683
50	54.296	59.059	55.761	56.834	57.587
70	77.521	84.142	79.586	81.064	82.157
100	110.976	120.121	113.891	115.904	117.518

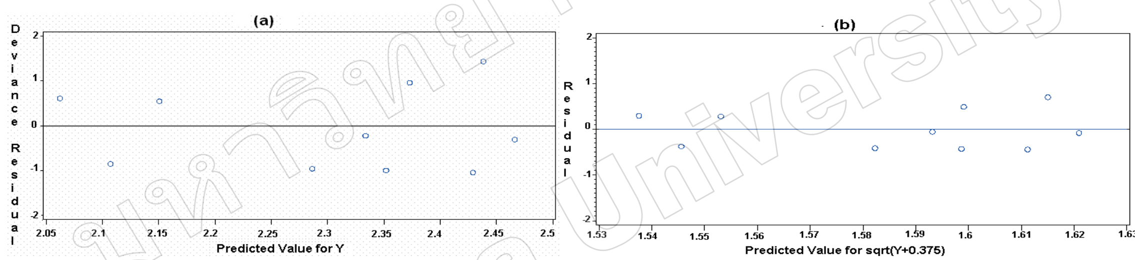


ภาพที่ 1 เปรียบเทียบค่า Deviance ที่เฉลี่ยจากการวนซ้ำ 500 รอบของตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y กับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลลักษณะ VST ในรูปแบบต่างๆ จำแนกตามขนาดตัวอย่าง

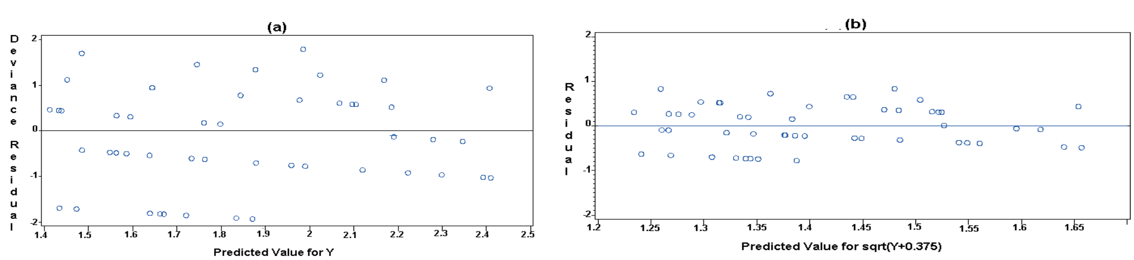
2 ผลการตรวจสอบตัวแบบถดถอยสำหรับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y โดยใช้วิธี Deviance Residual Plot และการตรวจสอบตัวแบบสำหรับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงในรูปแบบ $Y' = \sqrt{Y + 3/8}$ โดยใช้วิธี Residual Plot โดยงานวิจัยนี้จะแสดงผลการตรวจสอบตัวแบบในบางสถานการณ์จำลองเท่านั้นเนื่องจากสถานการณ์อื่นๆ ก็ให้ผลในทำนองเดียวกัน ดังนี้

ผลการตรวจสอบตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y โดยใช้วิธี Deviance Residual Plot กับค่าทำนายของ $Y(\hat{\mu})$ ในที่นี้จะได้แสดงผลเพียงที่ขนาดตัวอย่าง (n) เท่ากับ 10 และ 50 ดังแสดงในภาพที่ 2 (a) และภาพที่ 3 (a) ตามลำดับ เมื่อพิจารณาที่ขนาดตัวอย่างเท่ากับ 10 จากลักษณะการกระจายของค่า Deviance Residual ของตัวแบบพบว่าค่อนข้างกระจายห่างจากเส้นที่ค่า Deviance Residual เป็น 0 สำหรับเมื่อขนาดตัวอย่างมาก เช่น ในที่นี้นำเสนอกรณีขนาดตัวอย่างเท่ากับ 50 ก็ได้ผลทำนองเดียวกันนั่นหมายความว่าตัวแบบถดถอยปัวซงนี้มีค่าความแปรปรวนของความคลาดเคลื่อนค่อนข้างสูง

ส่วนผลการตรวจสอบตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ ในรูปแบบ $Y' = \sqrt{Y+3/8}$ โดยใช้วิธี Residual Plot กับค่าทำนายของ Y' ที่ขนาดตัวอย่างเท่ากับ 10 และ 50 ดังแสดงในภาพที่ 2 (b) และ 3 (b) ตามลำดับ จากรูปทั้งสองแสดงให้เห็นว่าความแปรปรวนของความเคลื่อนคลาดของตัวแบบมีค่าคงที่ และมีค่าเฉลี่ยของความคลาดเคลื่อนเป็น 0 คือค่าตกค้าง (Residual) มีการกระจายอย่างไม่มีระบบแบบแผนรอบๆ เส้น 0 และผลที่ได้คือค่าความแปรปรวนของความคลาดเคลื่อนของตัวแบบ หรือค่าความแปรปรวนของ $Y' = \sqrt{Y+3/8}$ มีค่าคงที่นั้นสอดคล้องกับการศึกษาของ Anscombe (1948), Bar-Lev and Enis (1988) และ Guan (2009) ที่การแปลงค่าข้อมูลปัวซงในรูปแบบ $Y' = \sqrt{Y+3/8}$ ซึ่งให้ค่าความแปรปรวนคงที่ โดยภาพที่ 2 (b) และ 3 (b) แสดงให้เห็นผลที่แตกต่างกันอย่างชัดเจน เมื่อเทียบกับการกระจายของค่า Deviance Residual ของตัวแบบถดถอยปัวซงที่ได้จากข้อมูลปัวซง Y ในภาพที่ 2 (a) และ 3 (a) นั่นคือตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงในรูปแบบ $Y' = \sqrt{Y+3/8}$ มีความเหมาะสมตามคุณสมบัติของตัวแบบถดถอยเชิงเส้น ดังนั้นจึงสามารถหาค่าประมาณพารามิเตอร์ในตัวแบบนี้ได้ด้วยวิธี OLS



ภาพที่ 2 ที่ขนาดตัวอย่าง (n) เท่ากับ 10 (a) Deviance Residual Plot กับค่าทำนายของ $Y(\hat{\mu})$; (b) Residual Plot กับค่าทำนายของ $Y' = \sqrt{Y+3/8}$

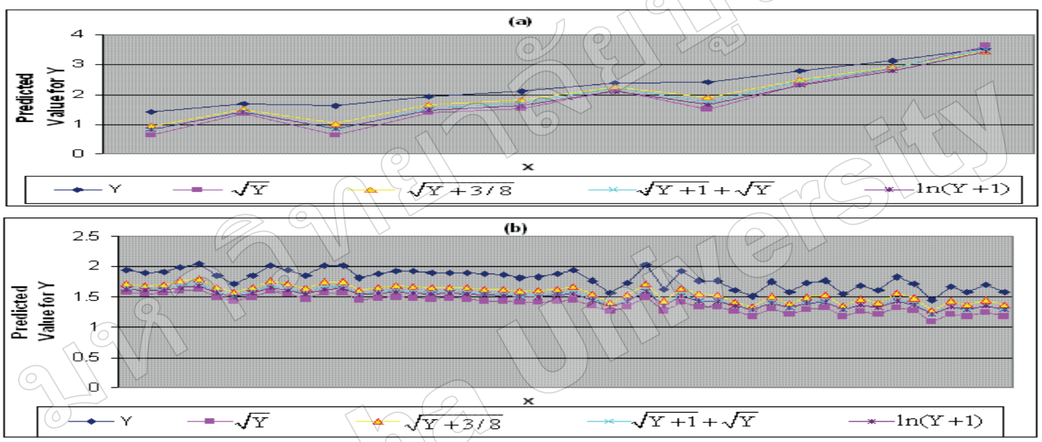


ภาพที่ 3 ที่ขนาดตัวอย่าง (n) เท่ากับ 50 (a) Deviance Residual Plot กับค่าทำนายของ $Y(\hat{\mu})$; (b) Residual Plot กับค่าทำนายของ $Y' = \sqrt{Y+3/8}$

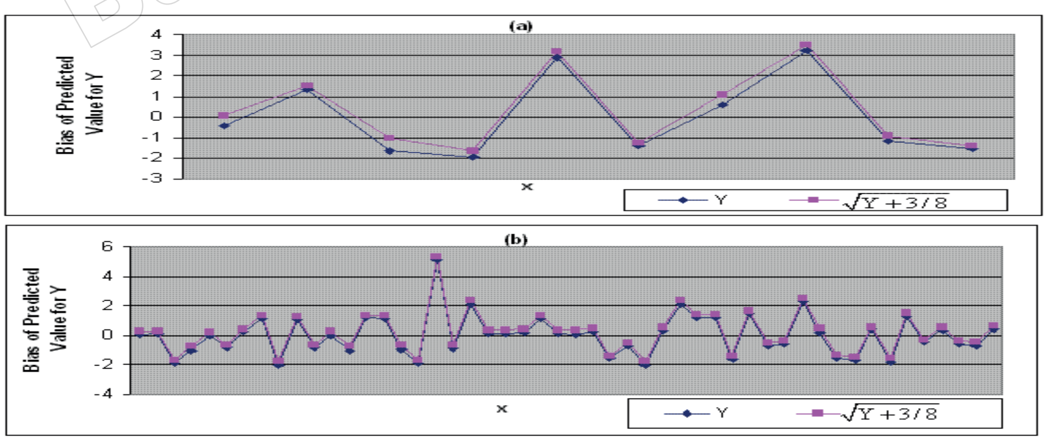
3 การเปรียบเทียบค่าทำนายของ $Y(\hat{\mu})$ จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ใน 4 รูปแบบ กับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y โดยตรง

ในสถานการณ์ที่ขนาดตัวอย่างเท่ากับ 10 ค่าทำนายของ $Y(\hat{\mu})$ จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงในรูปแบบ $Y' = \sqrt{Y+3/8}$ จะมีค่าเท่ากับ $\hat{\mu} = (\hat{Y}')^2 - 0.375$ เมื่อ $\hat{Y}' = \mathbf{x}^T\mathbf{b} = b_0 + b_1X_1 + b_2X_2$ ซึ่งค่าทำนายของ Y นี้จะให้ค่าใกล้เคียงกับค่าทำนายของ Y จากตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y มากที่สุด เมื่อเทียบกับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงด้วยรูปแบบอื่น ดังแสดงในภาพที่ 4 (a) และเมื่อขนาดตัวอย่างมากขึ้น เช่น 50 ผลการเปรียบเทียบค่าทำนายของ Y จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลในรูปแบบ $Y' = \sqrt{Y+3/8}$ ก็ยังให้ผลที่ใกล้เคียงมากที่สุดกับค่าทำนายของ Y จากตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y เมื่อเทียบกับวิธีการแปลงในรูปแบบอื่น ดังแสดงในภาพที่ 4 (b)

จากการเปรียบเทียบค่าทำนายของ Y ข้างต้น เมื่อนำผลการทำนายค่าของ Y ที่ได้จากทั้งตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบ $Y' = \sqrt{Y+3/8}$ และค่าทำนายค่า Y ที่ได้จากตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y มาพิจารณาเปรียบเทียบค่าความเอนเอียง (Bias) ของค่าทำนายที่แตกต่างไปจากข้อมูล Y ที่สถานการณ์ขนาดตัวอย่างเล็กๆ เท่ากับ 10 และขนาดตัวอย่างที่ใหญ่เท่ากับ 50 ดังภาพที่ 5 (a) และภาพที่ 5 (b) ตามลำดับ พบว่า เมื่อมีจำนวนตัวอย่างมาก จะยิ่งทำให้ค่าทำนายของ Y ที่ได้จากตัวแบบถดถอยทั้งสองตัวแบบมีค่าแทบจะไม่แตกต่างกัน



ภาพที่ 4 เปรียบเทียบค่าทำนายของ $Y(\hat{\mu})$ จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูล ลักษณะ VST ใน 4 รูปแบบกับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y (a) n=10; (b) n=50



ภาพที่ 5 เปรียบเทียบค่าเอนเอียง (Bias) ในค่าทำนายของ $Y(\hat{\mu})$ จากตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลลักษณะ VST รูปแบบ $\sqrt{Y+3/8}$ กับตัวแบบถดถอยปัวซงที่ได้จากข้อมูล Y (a) n=10; (b) n=50

สรุปผลและข้อเสนอแนะ

1. สรุปผล

ตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบ $Y' = \sqrt{Y+3/8}$ ที่ใช้การประมาณค่าพารามิเตอร์แบบ OLS จะมีลักษณะเป็นไปตามคุณสมบัติของตัวแบบถดถอยเชิงเส้น และมีความเหมาะสมกับข้อมูลแจกแจงแบบปัวซงมากที่สุด โดยให้ค่า Deviance ที่เฉลี่ยจากการวนซ้ำมีค่าต่ำสุดเมื่อเทียบกับตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลในรูปแบบอื่น โดยเฉพาะอย่างยิ่งเมื่อมีขนาดตัวอย่างใหญ่ขึ้น ยิ่งแสดงได้ชัดเจนถึงลักษณะของข้อมูลปัวซงที่ผ่านการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่มีคุณสมบัติของตัวแบบถดถอยเชิงเส้น คือมีลักษณะความแปรปรวนของความคลาดเคลื่อนคงที่ ค่าเฉลี่ยของค่าคลาดเคลื่อนเป็น 0 นอกจากนี้ค่าทำนายของ $Y(\hat{\mu})$ จากตัวแบบถดถอยเชิงเส้นนี้จะให้ค่าที่ใกล้เคียงกับค่าทำนายของ Y จากตัวแบบถดถอยปัวซงที่ได้จากข้อมูลปัวซงโดยตรง สรุปได้ว่าตัวแบบถดถอยเชิงเส้นที่ได้จากวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบ $Y' = \sqrt{Y+3/8}$ ที่ทำการประมาณค่าพารามิเตอร์ด้วยวิธี OLS จะมีความเหมาะสมที่จะใช้เป็นตัวแบบสำหรับข้อมูลที่มีการแจกแจงปัวซง

2. ข้อเสนอแนะ

การศึกษาวิจัยต่อไปควรพิจารณาตัวแปรทำนายมากกว่า 2 ตัวแปร ที่มีทั้งลักษณะค่าข้อมูลเป็นเชิงปริมาณ และเชิงคุณภาพ และเปรียบเทียบกับวิธีการแปลงข้อมูลเพื่อให้ได้ความแปรปรวนคงที่ในรูปแบบอื่นๆ ที่นอกเหนือจากการวิจัยนี้

เอกสารอ้างอิง

Anscombe, F.J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika 35*, 246-254. Available Source: http://en.wikipedia.org/wiki/Anscombe_transform.

Bar-Lev, S.K., and Enis, P. (1988). On the Classical Choice of Variance Stabilizing Transformations and an Application for a Poisson Variate. *Biometrika 75*, 803-804.

Freeman, M.F., and Tukey, J.W. (1950). Transformations Related to the Angular and the Square Root. *The Annual of Mathematical Statistical, 21*, 607-611.

Guan Yu. (2009). Variance Stabilizing Transformations of Poisson, Binomial and Negative Binomial Distributions. *Statistics and Probability Letters 79*, 1621-1629.

Kutner, M.H. , Nachtsheim, C.J., Neter,J., and Li,W. (2005). *Applied Linear Statistical Models*. Singapore: McGraw-Hill Companies,Inc.

Lin, S.M., Du, P., Huber, W., and Kibbe, W.A. (2008). Model-Based Variance-Stabilizing Transformation for Illumina Microaray Data. *Nucleic Acids Research 36*(2), e11, Available Source:http://nar.oxfordjournals.org/cgi/reprint/36/2/e11.pdf., Published Online 4 January 2008.

McCullagh, P., and Nelder, J.A. (1996). *Generalized Linear Models*. (2nd ed). London: Chapman and Hall.

Myers, R.H., and Milton, J.S. (1991). *A First Course in the Theory of* Linear Statistical Models. Boston: PWS-KENT Pub.Co.

Myers, R.H. ,Montgomery, D.C., and Vining, G.G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*. New York: John Wiley and Sons,Inc.

Rocke, D., and Durbin, B. (2003). Approximate Variance Stabilizing Transformation for Gene-Expression Microarray Data. Bioinformatics 19, 966-972. Available Source:http://bioinformatics.oxfordjournals.org/cgi/screenpdf/19/8/966.pdf.

Uddin, M.T., Noor, M.S., Kabir, A., Ali, R., and Islam, M.N. (2006). The Transformations of Random Variables under Symmetry. *Journal of Applied Sciences 6*(8),1818-1821.