บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันนี้งานวิจัยมีบทบาทสำคัญ หลายองค์กรมีการเก็บรวบรวมข้อมูลเพื่อนำมาทำงานวิจัย โดยใช้กระบวนการทางสถิติในการวิเคราะห์ประมวลผลและนำผลของการวิจัยปรับใช้กับแนวทางขององค์กร แต่ในบางครั้งการเก็บข้อมูลต้องใช้ระยะเวลานาน ไม่ว่าจะทางการแพทย์ การเกษตร ต้องเก็บข้อมูลจากหน่วยตัวอย่างเดียวกัน ซ้ำมากกว่า 1 ครั้ง ในระยะเวลาที่ต่างกัน เรียกว่า ข้อมูลระยะยาว

วิทยานิพนธ์เรื่องนี้เป็นการศึกษาการเปรียบเทียบประสิทธิภาพของตัวแบบ Generalized Linear Model และ ตัวแบบ Generalized Estimating Equations ด้วยข้อมูลระยะยาว โดยในการศึกษานี้จะจำลองสถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์กันและกำหนดอัตตสัมพันธ์และโครงสร้างความแปรปรวนร่วม เพื่อประมาณพารามิเตอร์ทั้งสองตัวแบบด้วยวิธีเดียวกันคือ วิธี Quasi-Likelihood แล้วเปรียบเทียบดูว่าตัวแบบใดเหมาะสมกว่าสำหรับข้อมูลระยะยาวที่มีตัวแปรตามมีการแจกจงแบบปัวซงส์ เพื่อสามารถนำไปใช้ประโยชน์ได้ในอนาคตต่อไป

การแจกแจงแบบปัวซงส์นั้นเป็นการแจกแจงปัวซงส์เป็นการแจกแจงที่เกิดขึ้นในตัวแปรสุ่มของการทดลองหรือว่าเหตุการณ์ที่เกิดขึ้นในช่วงระยะเวลาใดเวลาหนึ่งที่ต่อเนื่องกัน โดยหนึ่งหน่วยนั้นอาจจะเป็น วินาที นาที ชั่วโมง วัน สัปดาห์ เดือน หรือ ปี อาจจะเป็นความยาว พื้นที่ ปริมาตร โดยข้อมูลที่ได้จากตัวแปรสุ่มนี้เป็นแบบไม่ต่อเนื่องกันชนิดหนึ่ง ตัวอย่างของตัวแปรตามที่มีการแจกแจงแบบปัวซงส์พบอยู่ทั่วไป เช่น อัตราการเกิดของทารกที่เกิดขึ้นในหนึ่งปีของกรุงเทพมหานคร จำนวนครั้งที่มีการกดเอทีเอ็มในแต่ละวัน เป็นต้น เราสามารถเขียนฟังก์ชั่นการแจกแจงในรูปนี้ $f(Y_i)=\frac{(t_i\mu_i)^{Y_i}\exp(-t_i\mu_i)}{Y_i!};Y_i=0,1,2,3,...$ โดยที่ t_i คือ ระยะเวลาในการเก็บข้อมูลซ้ำ ตัวแบบ Generalized Estimating Equationsเป็นตัวแบบที่ขยายจากตัวแบบ Generalized Linear Model ภายใต้สถานการณ์ที่ข้อมูลมีความสัมพันธ์กัน สามารถเก็บข้อมูลซ้ำได้หลายช่วงเวลา ซึ่งเป็นตัวแบบที่นิยมมากที่ใช้กับข้อมูลที่เป็นกลุ่มและข้อมูลการนับ โดยที่ตัวแปรตามไม่จำเป็นต้องมีการแจกแจงแบบปกติ สามารถมีการแจกแจงแบบ Binomial , Poisson , Gamma เป็นต้น สามารถเขียนในรูปแบบได้ดังนี้

$$\log_e(Y_{i(t)})=\beta_0+\beta_1 t+\beta_{2j}\sum_{j=1}^{J}x_{ijt}+corr+e_{i(t)}$$

$$i=1,2,\ldots,n \qquad j=1,2,\ldots,J \qquad t=1,2,\ldots,T$$

และตัวแบบ Generalized Linear Model ที่นำมาใช้ในการเปรียบเทียบนี้ คือ Poisson Regression ซึ่งมีตัวแปรตามมีการแจกแจงแบบปัวซงส์เช่นกัน สามารถเขียนในรูปแบบดังนี้

$$\log_e(Y_{i(t)}) = \beta_0 + \beta_1 t + \sum_{j=1}^{J} \beta_{2j} x_{ijt} + \varepsilon_{i(t)} \quad i = 1,2,\ldots,n \qquad j = 1,2,\ldots,J$$

$t = 1,2,\ldots,T$ ซึ่งรายละเอียดของตัวแปรต่างๆในสมการจะอธิบายในบทถัดไป วิธีที่ใช้ในการประมาณค่าพารามิเตอร์นี้ จะเลือกใช้วิธี Quasi – Likelihood ซึ่งเป็นการประมาณที่ดึงมาส่วนหนึ่งของการประมาณMaximum Likelihood เนื่องจากสามารถใช้ประมาณค่าพารามิเตอร์ได้ทั้งสองตัวแบบ ซึ่งมีรูปแบบที่สนใจดังนี้ $U(\beta) = 0$ ซึ่งจะได้

$$U(\beta) = \sum D^T V^{-1}(Y - \mu) = 0$$

ซึ่งเรียกว่า quasi – score function โดยที่

$$D_i = \frac{\partial \mu_i}{\partial \beta_t}$$

$$V_i = (A_i^{1/2} R_i A_i^{1/2})\phi$$

$$\phi = \frac{1}{n-p} \sum_i \sum_j \frac{y_{it} - \mu_{it}}{\sqrt{\mathrm{var}(\mu_{it})}}$$

เมื่อ V_i คือ เมทริกซ์โครงสร้างความแปรปรวนร่วมของ Y_{it}

A_i คือ diagonal matrix ความแปรปรวนของ Y_{it}

R_i คือ เมทริกซ์อัตตสัมพันธ์ของ Y_{it}

ϕ คือ overdispersion parameter

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาเปรียบเทียบประสิทธิภาพ ของตัวแบบ Generalized Linear Model และ ตัวแบบ Generalized Estimating Equations ด้วยวิธีการประมาณพารามิเตอร์แบบ Quasi-Likelihood สำหรับข้อมูลระยะยาว เมื่อกำหนดให้ตัวแปรตามมีการแจกแจงปัวซงส์

ขอบเขตของการวิจัย

ในการวิจัยครั้งนี้กระทำภายใต้ขอบเขตดังนี้

1. วิธีการประมาณค่าพารามิเตอร์ที่สนใจในการศึกษาในงานวิจัย คือ Quasi – Likelihood ซึ่งมีรูปแบบดังนี้

$$U = u(\mu; Y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

$$Q(\mu;y)=\int_{y}^{\mu}u(y|y)dt=\int_{y}^{\mu}\frac{y-t}{\sigma^2V(t)}dt$$

$$U(\beta)\ =\ \sum D^TV^{-1}(Y-\mu)\ =\ 0$$

2. จำนวนตัวแปรอิสระที่นำมาศึกษาเท่ากับ 1 , 3
3. ขนาดตัวอย่างที่ศึกษาเท่ากับ 20 , 60
4. กำหนดให้อัตตสัมพันธ์และโครงสร้างความแปรปรวนร่วมของตัวแบบ Generalized Linear Model และ ตัวแบบ Generalized Estimating Equations คือ Exchangeable
5. กำหนดรูปแบบความสัมพันธ์ระหว่างตัวแปรอิสระด้วยกัน คือ มีสัมพันธ์กันเล็กน้อย สัมพันธ์กันปานกลาง และสัมพันธ์กันมาก คือ 0.1 , 0.5 ,0.9
6. ศึกษาเมื่อระยะเวลาที่ทำการเก็บข้อมูลซ้ำ (t) เท่ากับ 3 , 6
7. กำหนดการสุ่มตัวอย่างจำลองสถานการณ์ 1000 รอบ
8. เปรียบเทียบตัวแบบด้วยค่าเฉลี่ยของผลบวกกำลังสองของความคลาดเคลื่อนของสัมประสิทธิ์ความถดถอย (AMSE)

ข้อตกลงเบื้องต้น

ในการวิเคราะห์ครั้งนี้ผู้วิจัยได้กำหนดข้อตกลงเบื้องต้นดังนี้

1. ตัวแปรตามมีการแจกแจงแบบปัวซงส์
2. ตัวแปรอิสระเป็นเชิงปริมาณและมีความสัมพันธ์กัน
3. ข้อมูลที่ศึกษาไม่มีค่าติดลบ

เกณฑ์ที่ใช้ในการตัดสินใจ

ในการวิจัยนี้จะพิจารณาค่าเฉลี่ยของผลบวกกำลังสองของความคลาดเคลื่อนของสัมประสิทธิ์ความถดถอย (Average Mean Square Error : AMSE) ซึ่งมีสูตรการคำนวณดังนี้

$$AMSE=\frac{\sum_{k=1}^{1000}\sum_{m=0}^{M}\left(\hat{\beta}_{mk}-\beta_{mk}\right)^2}{1000M}\qquad \text{โดยที่}\quad m=0,1,\ldots,M\quad k=1,2,\ldots,1000$$

$\hat{\beta}_{mk}$ คือ ค่าประมาณของพารามิเตอร์ตัวที่ m ครั้งที่ k

β_{mk} คือ ค่าจริงของพารามิเตอร์ตัวที่ m ครั้งที่ k

M คือ จำนวนของพารามิเตอร์ที่มีอยู่ในสมการ

คำจำกัดความที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้มีคำจำกัดความที่ใช้ดังนี้

1. ข้อมูลระยะยาว (Longitudinal Data) หมายถึง ข้อมูลที่มีการเก็บรวบรวมจากหน่วยทดลองเดิมมากกว่า 1 ครั้งขึ้นไปในช่วงเวลาที่ต่างกัน
2. พารามิเตอร์ (Parameter) หมายถึง ค่าคงที่ที่แสดงคุณลักษณะบางประการของประชากร
3. Exchangeable หมายถึง ความสัมพันธ์ของค่าสังเกตภายในหน่วยศึกษาเดียวกันจะมีค่าเท่ากัน ณ ช่วงเวลาที่ต่างกัน

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อทราบถึงตัวแบบที่เหมาะสมในการวิเคราะห์ข้อมูล
2. เพื่อทราบถึงวิธีการประมาณค่าพารามิเตอร์ที่ตัวแปรตามมีการแจกแจงแบบปัวซงส์ด้วยวิธี Quasi - Likelihood
3. เพื่อเป็นแนวทางในการศึกษาวิจัยต่อไป

วิธีดำเนินการวิจัย

1. สร้างตัวแปรตามที่มีการแจกแจงแบบปัวซงส์
2. สร้างตัวแปรอิสระโดยกำหนดให้เป็นค่าคงที่ โดยกำหนดให้มีค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ของตัวแปรอิสระ
3. สร้างโครงสัมพันธ์ของตัวแปรตามของตัวแบบ Generalized Estimating Equations แบบ Exchangeable
4. สร้างโครงสร้างความแปรปรวนร่วมของแบบ Generalized Linear Model แบบ Exchangeable

5. ทำการประมาณพารามิเตอร์ของตัวแบบ Generalized Linear Model ด้วยวิธี Quasi-Likelihood
6. ทำการประมาณพารามิเตอร์ของตัวแบบ Generalized Estimating Equations ด้วยวิธี Quasi-Likelihood
7. ทำการเปรียบเทียบวิธีการประมาณตัวแบบ Generalized Linear Model และ ตัวแบบ Generalized Estimating Equations ที่ได้การวิธี Quasi-Likelihood โดยเปรียบเทียบด้วยวิธี ค่าเฉลี่ยของผลบวกกำลังสองของความคลาดเคลื่อน (Average Mean Square Error : AMSE)
8. สรุปผลการวิจัยในแต่ละสถานการณ์