การเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหาย ในการถดถอยปัวซง

ศุภรัสมิ์ แย้มมีเชาว์

วิทยาศาสตรมหาบัณฑิต

สาขาวิชาสถิติประยุกต์

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บัณฑิตวิทยาลัย

มหาวิทยาลัยเชียงใหม่

ธันวาคม 2561

การเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหาย ในการถดถอยปัวซง

ศุภรัสมิ์ แย้มมีเชาว์

การค้นคว้าแบบอิสระนี้เสนอต่อมหาวิทยาลัยเชียงใหม่เพื่อเป็นส่วนหนึ่งของ
การศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่
ธันวาคม 2561

การเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหาย ในการถดถอยปัวซง

ศุภรัสมิ์ แย้มมีเชาว์

การค้นคว้าแบบอิสระนี้ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์

คณะกรรมการสอบ

.....ประธานกรรมการ
(ดร.ชูเกียรติ ผุดพรมราช)

.....กรรมการ
(รศ. พุฒิพงษ์ พุกกะมาน)

.....กรรมการ
(ผศ.ดร. บัณฑิตา พลับอินทร์)

.....กรรมการ
(ผศ.ดร. มานัดถุ์ คำกอง)

คณะกรรมการที่ปรึกษา

..... อาจารย์ที่ปรึกษาหลัก
(รศ. พุฒิพงษ์ พุกกะมาน)

..... อาจารย์ที่ปรึกษาร่วม
(ผศ.ดร. บัณฑิตา พลับอินทร์)

.....อาจารย์ที่ปรึกษาร่วม
(ผศ.ดร. มานัดถุ์ คำกอง)

25 ธันวาคม 2561

© ลิขสิทธิ์ของมหาวิทยาลัยเชียงใหม่

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

กิตติกรรมประกาศ

การค้นคว้าแบบอิสระฉบับนี้สำเร็จลุล่วงไปด้วยดีตามความคาดหวังของผู้วิจัยได้ เนื่องจากความเมตตากรุณาจากท่านอาจารย์ รศ. พุฒิพงษ์ พุกกะมาน อาจารย์ที่ปรึกษาการค้นคว้าแบบอิสระ ที่ท่านได้เสียสละเวลาที่มีค่าในการให้คำปรึกษาเกี่ยวกับแนวคิดในการทำการค้นคว้าแบบอิสระตลอดจนคำแนะนำทางวิชาการ รวมถึงตรวจสอบแก้ไขข้อบกพร่องและความเรียบร้อยของการค้นคว้าแบบอิสระเล่มนี้จนกระทั่งเสร็จสมบูรณ์

ขอขอบพระคุณ ดร. ชูเกียรติ ผุดพรมราช ประธานกรรมการสอบการค้นคว้าแบบอิสระ ที่กรุณาให้คำปรึกษาและข้อเสนอแนะเพิ่มเติม ตลอดจนแก้ไขข้อบกพร่อง ในการทำการค้นคว้าแบบอิสระให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณ ผศ.ดร. บัณฑิตา พลับอินทร์ และผศ.ดร. มานัดถุ์ คำกอง อาจารย์ที่ปรึกษาร่วมและกรรมการสอบการค้นคว้าแบบอิสระ ที่กรุณาให้คำปรึกษาแนะนำและพิจารณาตรวจสอบแก้ไขการค้นคว้าแบบอิสระให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ได้ถ่ายทอดวิชาความรู้ด้านสถิติอันเป็นประโยชน์ในการศึกษา และขอขอบพระคุณเจ้าหน้าที่ภาควิชาสถิติที่ได้อำนวยความสะดวกในการทำการค้นคว้าอิสระครั้งนี้ ทั้งยังช่วยในเรื่องการประสานงานต่างๆ และขอขอบคุณเพื่อนๆ นักศึกษาปริญญาโท สาขาสถิติประยุกต์ทุกท่านที่เป็นกำลังใจและให้ความช่วยเหลือในการทำการค้นคว้าแบบอิสระเรื่องนี้เป็นอย่างดีตลอดมา

สุดท้ายนี้ ผู้วิจัยขอขอบพระคุณ บิดา มารดา และครอบครัว ซึ่งเป็นผู้สนับสนุนและให้โอกาสทางการศึกษาเล่าเรียน ตลอดจนคอยช่วยเหลือและให้กำลังใจผู้วิจัยเสมอมากระทั่งสำเร็จการศึกษา

ศุภรัสมิ์ แย้มมีเชาว์

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

หัวข้อการค้นคว้าแบบอิสระ	การเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายในการถดถอยป๊วซง
ผู้เขียน	นางสาวศุภรัสมิ์ แย้มมีเชาว์
ปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
คณะกรรมการที่ปรึกษา	รศ.พุฒิพงษ์ พุกกะมาน อาจารย์ที่ปรึกษาหลัก ผศ.ดร.บัณฑิตา พลับอินทร์ อาจารย์ที่ปรึกษาร่วม ผศ.ดร.มานัดถุ์ คำกอง อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายสามวิธีในการถดถอยแบบป๊วซง ได้แก่ วิธีระยะห่างใกล้สุด (Nearest Neighbor Imputation: NNI) วิธีการถดถอยป๊วซง (Poisson Regression Imputation: PRI) และวิธีระยะห่างใกล้สุดและวิธีการถดถอยป๊วซงถ่วงน้ำหนัก (Weighted Nearest Neighbor and Poisson Regression: WNPR) เพื่อเสนอวิธีที่ดีที่สุดในการจัดการกับปัญหาค่าสูญหายในการถดถอยแบบป๊วซง โดยใช้การจำลองข้อมูลแบบมอนติคาร์โลโดยมีตัวแปรอิสระสามตัวแปรที่มีการแจกแจงเอกรูป ตัวแปรตามหนึ่งตัวแปรมีการแจกแจงป๊วซง ขนาดตัวอย่างเป็น 20, 60, 100 และ 200 กำหนดให้มีการสูญหายอย่างสุ่มเฉพาะบางส่วน (Missing At Random : MAR) ร้อยละการสูญหายเป็น 5, 10 และ15 โดยกำหนดค่าพารามิเตอร์ดังต่อไปนี้

ชุดที่ 1 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$

ชุดที่2 คือ $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$

ชุดที่ 3 คือ $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$

ชุดที่ 4 คือ $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$

และชุดที่ 5 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$

เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการที่เหมาะสมคือค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าน้อยที่สุด (Mean Square Error: MSE)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ผลการศึกษาพบว่าในพารามิเตอร์ชุดที่ 1 วิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดพารามิเตอร์ชุดที่ 2 พบว่าวิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุดที่เมื่อขนาดตัวอย่างเป็น 100 และ 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 20 และ 60พารามิเตอร์ชุดที่ 3 พบว่าวิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุดที่ในกรณีที่ขนาดตัวอย่าง 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่าง 20, 60 และ 100พารามิเตอร์ชุดที่ 4 พบว่าวิธีการถดถอยป๊วซงให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด รองลงมาคือวิธีถ่วงน้ำหนักและวิธีระยะห่างใกล้สุดตามลำดับ วิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุดและพารามิเตอร์ชุดที่ 5 พบว่าวิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 60, 100 และ 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 20

คำสำคัญ: ค่าสูญหาย การประมาณค่า ค่าพารามิเตอร์ การถดถอยป๊วซง และตัวแปรจำนวนนับ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Independent Study Title	Comparison of the Estimation Methods for Missing Data in Poisson Regression
Author	Miss Suppharat Yaemmeechao
Degree	Master of Science (Applied Statistics)
Advisory Committee	Assoc. Prof. Putipong Bookkamana Advisor Asst. Prof. Dr. Bandhita Plubin Co-advisor Asst. Prof. Dr. Manad Khamkong Co-advisor

ABSTRACT

The objective of this study is to compare the three methods of the missing values in Poisson Regression model; the Nearest Neighbor Imputation (NNI), the Poisson Regression Imputation (PRI) and the Weighted Nearest Neighbor and Poisson Regression (WNPR) in order to propose the best method to deal with missing value issues for the Poisson regression model. Using Monte Carlo simulation, three independent variables with Uniform distribution, and dependent variable with Poisson distribution were generated. The sample size is 20, 60, 100 and 200, requiring a Missing at Random (MAR). The percentage of missing value is 5, 10 and 15 with the following parameters.

Parameters set 1: $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ and $\beta_3 = 0.5$

Parameters set 2: $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ and $\beta_3 = 0.7$

Parameters set 3: $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ and $\beta_3 = -0.2$

Parameters set 4: $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ and $\beta_3 = -0.7$

Parameters set 5: $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ and $\beta_3 = 0.5$

The criteria for comparing the most suitable method is Mean Square Error (MSE)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

The results showed that Parameters set 1; the WNPR is the most appropriate method. The parameters set 2 showed that the PRI is the most appropriate method when the sample size is 100 and 200 while the WNPR is the most suitable method when the sample size is 20 and 60. The Parameters set 3 showed that the PRI is the most appropriate method when the sample size is 200 while the WNPR is the most appropriate method when the sample size is 20, 60 and 100. The Parameters set 4 showed that the PRI gave the lowest mean square error, followed by WNPR and NNI, respectively thus PRI is the most appropriate method. And the parameters set 5 showed that PRI is the most appropriate method when the sample size is 60, 100 and 200 while the WNPR is the most suitable method when the sample size is 20.

Keywords: Missing Value, Estimation, Parameter Value, Poisson Regression and Count Data

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

สารบัญ

หน้า

กิตติกรรมประกาศ ค

บทคัดย่อภาษาไทย ง

บทคัดย่อภาษาอังกฤษ ฉ

สารบัญตาราง ญ

สารบัญภาพ ฎ

บทที่ 1 บทนำ 1

1.1 ประวัติความเป็นมา 1

1.2 วัตถุประสงค์ของการศึกษา 3

1.3 ขอบเขตของงานวิจัย 3

1.4 ประโยชน์ที่จะได้รับจากการศึกษา 4

1.5 นิยามศัพท์ 4

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 6

2.1 แนวคิดและทฤษฎี 6

2.1.1 การแจกแจงที่เกี่ยวข้อง 6

2.1.2 ตัวแบบการถดถอยที่เกี่ยวข้อง 7

2.1.3 ข้อมูลสูญหาย (Missing Data) 9

2.1.4 ประเภทของข้อมูลสูญหาย (Type of Missing Data) 9

2.1.5 การประมาณค่าสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) 10

2.1.5 เกณฑ์ในการตัดสินใจ 12

2.2 งานวิจัยที่เกี่ยวข้อง 12

บทที่ 3 วิธีการดำเนินการวิจัย 16

3.1 แผนการดำเนินการวิจัย 16

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.2 ขั้นตอนการดำเนินการวิจัย 16

บทที่ 4 ผลการศึกษา 17

4.1 ผลการประมาณค่าสูญหายของข้อมูลด้วยวิธีการประมาณค่าทั้ง 3 วิธี ในแต่ละสถานการณ์ 20

บทที่ 5 สรุปและอภิปรายผล 31

5.1 สรุปผลการศึกษา 31

5.1.1 ผลการศึกษาในการเปรียบเทียบค่าคลาดเคลื่อนกำลังสองเฉลี่ย 31

5.2 อภิปรายผลการศึกษา 33

5.3 ข้อเสนอแนะ 34

เอกสารอ้างอิง 35

รายการสิ่งเผยแพร่ 36

ภาคผนวก 37

ประวัติผู้เขียน 44

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

สารบัญตาราง

หน้า

ตารางที่ 2.1 แสดงตัวแปรอิสระและตัวแปรแบบสมบูรณ์และแบบที่มีการสูญหาย 10

ตารางที่ 4.1 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหาย 3 วิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$ 21

ตารางที่ 4.2 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหาย 3 วิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์ $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$ 22

ตารางที่ 4.3 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหาย 3 วิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์ $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$ 23

ตารางที่ 4.4 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหาย 3 วิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์ $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$ 24

ตารางที่ 4.5 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหาย 3 วิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$ 25

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

สารบัญภาพ

		หน้า
ภาพที่ 3.1	ขั้นตอนการทำงานในการประมาณค่าพารามิเตอร์โดยใช้ข้อมูลจำลอง	18
ภาพที่ 4.1	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI PRI และ WMPR ที่มีค่าพารามิเตอร์เป็น $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$	26
ภาพที่ 4.2	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI PRI และ WMPR ที่มีค่าพารามิเตอร์เป็น $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$	27
ภาพที่ 4.3	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI PRI และ WMPR ที่มีค่าพารามิเตอร์เป็น $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$	28
ภาพที่ 4.4	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI PRI และ WMPR ที่มีค่าพารามิเตอร์เป็น $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$	29
ภาพที่ 4.5	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI PRI และ WMPR ที่มีค่าพารามิเตอร์เป็น $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$	30

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บทที่ 1

บทนำ

1.1 ประวัติความเป็นมา

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีทางสถิติที่ใช้หาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยที่ตัวแปรตามเป็นเชิงปริมาณ เพื่อใช้ในการพยากรณ์ค่าของตัวแปรหนึ่งจากตัวแปรอื่นๆ การวิเคราะห์การถดถอยระหว่างตัวแปรอิสระหนึ่งตัวและตัวแปรตามหนึ่งตัว โดยที่ตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นตรงกันเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis) หากมีตัวแปรอิสระมากกว่าหนึ่งตัว และตัวแปรตามเพียงตัวเดียวเรียกว่า การวิเคราะห์การถดถอยพหุ (Multiple Regression Analysis) (พรสิน สุภวาลย์, 2556) แต่ในการวิเคราะห์ข้อมูลจากตัวอย่างที่เก็บรวบรวมมาได้นั้น เมื่อตัวแปรตอบสนองหรือตัวแปรตามเป็นจำนวนนับ การใช้การวิเคราะห์การถดถอยทั่วไปทำให้การเกิดความคลาดเคลื่อนสูง นักวิจัยมักนิยมใช้การวิเคราะห์การถดถอยป๊วซง (Poisson Regression Analysis) ในการจัดการกับข้อมูลดังกล่าว เพื่อทำนายหรือพยากรณ์ค่าของตัวแปรตาม เช่น ปัจจัยที่ส่งผลต่อจำนวนครั้งของครอบครัวที่จะไปตลาดประจำหมู่บ้านใน 1สัปดาห์ โดยพิจารณาจากปัจจัยต่างๆ โดยที่ตัวแปรที่ศึกษาหรือตัวแปรตามเป็นตัวแปรไม่ต่อเนื่องที่มีการแจกแจงป๊วซง และตัวแปรที่แทนปัจจัยต่างๆ หรือตัวแปรอิสระ เป็นตัวแปรเชิงปริมาณ หรืออาจมีตัวแปรเชิงกลุ่มอยู่ด้วยและต้องปรับตัวแปรเชิงกลุ่มให้เป็นตัวแปรหุ่น (ชูเกียรติ ผุดพรมราช, 2547)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ในการวิเคราะห์ข้อมูล การพยากรณ์ หรืองานวิจัย ปัญหาที่พบในการวิเคราะห์ทางสถิติโดยทั่วไปปัญหาหนึ่งคือข้อมูลสูญหาย (Missing Data) ข้อมูลสูญหายเป็นข้อมูลที่ไม่ทราบค่า หากนำข้อมูลเหล่านี้มาวิเคราะห์จะทำให้เกิดความเอนเอียง หรือเกิดความคลาดเคลื่อนสูงได้ ก่อนที่จะนำข้อมูลเหล่านี้มาวิเคราะห์ต้องทำการจัดการกับข้อมูลเสียก่อนเพื่อให้ได้ข้อมูลที่สมบูรณ์และเกิดความคลาดเคลื่อนน้อยที่สุด การจัดการกับข้อมูลสูญหาย นรุตม์ บุตรพลอย (2553) กล่าวว่านักวิจัยได้พยายามที่จะจัดการข้อมูลที่มีค่าสูญหาย เทคนิคที่จะนำมาประมาณค่าสูญหายมีหลายวิธี เช่น การหาตัวแทนของข้อมูล หรือการแทนค่าสูญหายด้วยค่ากลาง (Mean) หรือค่าฐานนิยม (Mode) ของกลุ่ม

นั้นๆ ซึ่งเป็นวิธีที่ง่ายแต่อาจทำให้เกิดการเบี่ยงเบนของข้อมูลดังนั้นจึงมีนักวิจัยได้พัฒนาและศึกษาเพื่อหาวิธีประมาณค่าสูญหายของข้อมูล โดยสามารถแบ่งเทคนิคออกเป็น 3 ประเภท ดังนี้

1. การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (Ignoring and Discarding Data) วิธีการนี้แบ่งเป็นสองวิธีดังนี้

    1.1 Listwise Deletion เป็นวิธีตัดแถวที่มีค่าข้อมูลสูญหายทิ้งทั้งหมด ซึ่งเป็นวิธีที่ง่ายที่สุดเพื่อให้ได้มาซึ่งกลุ่มข้อมูลที่สมบูรณ์ที่สุด

    1.2 Pairwise Deletion จะไม่ตัดแถวที่ขาดความสมบูรณ์ทิ้งไปแต่จะนำแถวเหล่านั้นมาใช้ในการประมวลผลด้วย จะเป็นการพิจารณาทุกแถวที่มีค่าที่สนใจ

2. การประมาณค่าพารามิเตอร์ (Parameter Estimation) ตัวอย่างวิธีการประมาณค่าได้แก่ การประมาณค่าความเป็นไปได้มากที่สุด (Maximum Likelihood Estimation) ซึ่งคำนวณจากข้อมูลที่ทราบทั้งหมด ค่าความเป็นไปได้มากที่สุดจะถูกคำนวณโดยแยกคำนวณจากข้อมูลที่สมบูรณ์ของบางตัวแปรและคำนวณจากข้อมูลในทุกตัวแปร ค่าความเป็นไปได้มากที่สุดจากการคำนวณทั้งสอง จะถูกใช้ในการประมาณค่าสูญหายอีกทีหนึ่ง
3. การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) วิธีการนี้จะทำการประมาณค่าข้อมูลสูญหายโดยอาศัยความสัมพันธ์ระหว่างกลุ่มข้อมูล เพื่อนำไปประมาณค่าข้อมูลสูญหาย การประมาณค่าสูญหายจากการคำนวณ (Imputation) แบ่งออกเป็นสองกลุ่มหลัก (Laaksonen, 2000) ได้แก่

    3.1 Model-donor Imputation คือ การประมาณค่าที่ได้มาจากตัวแบบ (Model) ซึ่งมีอยู่หลากหลายวิธี เช่น Mean Imputation, Regression Imputation, Ratio Imputation และ Multiple Imputation

    3.2 Real-donor Imputation คือการประมาณค่าที่ได้จากเซตข้อมูลของค่าที่สังเกตได้ วิธีที่ใช้กันทั่วไป เช่น Cold Deck Imputation, Hot Deck Imputation และ Nearest Neighbor Imputation

กรณีตัวแปรตามมีลักษณะเป็นจำนวนนับที่มีข้อมูลสูญหาย จะมีวิธีการประมาณค่าสูญหายได้แก่ วิธีการประมาณค่าสูญหายแบบวิธีระห่างใกล้สุด Nearest Neighbor Imputation (NNI) ซึ่งเป็นวิธีการที่ตัวแปรอิสระที่สมบูรณ์มาคำนวณเพื่อหาระยะห่าง วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI) เป็นวิธีที่นำแนวคิดจากวิธี Regression Imputation ถูกจำลองภายใต้การถดถอยปัวซง และวิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Regression (WNPR) ซึ่งเป็นนำแนวคิดของทั้งสองวิธี คือ วิธีระยะห่างใกล้สุด Nearest Neighbor Imputation (NNI) และวิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI) มาคำนวณค่าข้อมูลสูญหายโดยการถ่วงน้ำหนัก

ดังนั้นในการศึกษาครั้งนี้ ผู้วิจัยสนใจที่จะศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายด้วย 1. วิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation (NNI) 2. วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI) 3. วิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson Regression (WNPR) และใช้เกณฑ์ในการตัดสินใจว่าการประมาณค่าสูญหายด้วยวิธีใดใช้ได้ดีกว่า โดยจะเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงในรูปแบบ Mean Square Error (MSE)

1.2 วัตถุประสงค์ของการศึกษา

เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation (NNI) วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI) และวิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson Regression (WNPR)

1.3 ขอบเขตของงานวิจัย

ผู้วิจัยได้ศึกษาในประชากรภายใต้การถดถอยปัวซง ซึ่งเป็นการจำลองข้อมูลโดยใช้โปรแกรม R เวอร์ชั่น 3.3.1 โดยผู้วิจัยได้กำหนดขอบเขตดังนี้

1.3.1 ประชากร ในการศึกษาครั้งนี้ผู้วิจัยได้กำหนดให้ข้อมูลตัวแปรตามมีการแจกแจงปัวซง

1.3.2 ตัวอย่าง ในการศึกษาครั้งนี้ผู้วิจัยตัวอย่าง คือ ข้อมูลที่สุ่มจากประชากรที่ที่มีการแจกแจงปัวซง ที่มีพารามิเตอร์ λ โดยที่ $\lambda = e^{X\beta}$ ให้มีขนาดตัวอย่าง 20, 60, 100 และ 200 หน่วย

1.3.3 ตัวแปร ประกอบไปด้วยตัวแปรอิสระ X สามตัวแปร ที่มีการแจกแจงเอกรูป U(0,1) และตัวแปรตาม Y หนึ่งตัวแปร

1.3.4 พารามิเตอร์ ประกอบไปด้วย ความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับตัวแปรตามทั้งหมด 3 ชุด

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ชุดที่ 1 คือ $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ -0.2 \\ 0.5 \end{bmatrix}$

ชุดที่ 2 คือ $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.2 \\ 0.2 \\ 0.7 \end{bmatrix}$

ชุดที่ 3 คือ $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.7 \\ -0.2 \end{bmatrix}$

ชุดที่ 4 คือ $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.8 \\ -0.2 \\ -0.7 \end{bmatrix}$

ชุดที่ 5 คือ $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.8 \\ 0.5 \end{bmatrix}$

ในแต่ละสถานการณ์มีการทำซ้ำ 1,000 ครั้ง โดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์

1.4 ประโยชน์ที่จะได้รับจากการศึกษา

เป็นแนวทางในการประมาณค่าสูญหายวิธีต่างๆ ภายใต้การถดถอยป๊วซง เพื่อนำไปใช้ในการวิเคราะห์ข้อมูลต่อไป

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

1.5 นิยามศัพท์

1.5.1 ค่าสูญหาย คือ ค่าที่ไม่สามารถทราบได้จากตัวแปรตาม หรือค่าที่เกิดการสูญหายไป

1.5.2 การประมาณค่า คือ การประมาณหรือการแทนของที่สนใจศึกษาในประชากรโดยไม่ทราบค่าของตัวแปรตามว่าเป็นเท่าใด

1.5.3 ค่าพารามิเตอร์ คือ ตัวเลขหรือค่าคงที่ที่แสดงลักษณะบางประการของประชากร หรือค่าที่คำนวณได้จากประชากร

1.5.4 การถดถอยปัวซง คือ ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตอบสนองหนึ่งตัวแปรโดยที่ตัวแปรตามเป็นจำนวนนับที่มีการแจกแจงปัวซง

1.5.5 ตัวแปรจำนวนนับ คือ ตัวแปรตามที่มีลักษณะเป็นจำนวนเต็มที่สามารถนับจำนวนได้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งศึกษาเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายในการถดถอยปัวซง (Comparison of the Estimation Methods for Missing Data in Poisson Regression) ในบทนี้จะกล่าวถึงแนวคิดทฤษฎี รูปแบบของข้อมูลที่สูญหายที่จำลองขึ้นภายใต้การแจกแจงปัวซง เทคนิคที่ใช้ในการประมาณค่าข้อมูลสูญหาย และการเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหาย เกณฑ์ที่ให้ในตัดสินใจ รวมทั้งการกล่าวถึงงานวิจัยต่างๆ ที่เกี่ยวข้อง รวบรวมและพัฒนาเพื่อการแก้ไขปัญหาค่าสูญหาย

2.1 แนวคิดและทฤษฎี

2.1.1 การแจกแจงที่เกี่ยวข้อง

1) การแจกแจงปัวซง

การแจกแจงปัวซง (Poisson distribution) เป็นการแจกแจงซึ่งอธิบายถึงจำนวนครั้งของเหตุการณ์หรือจำนวนสิ่งที่สนใจที่เกิดขึ้นในช่วงเวลาที่กำหนดหรือขอบเขตที่กำหนด เช่น จำนวนผู้เสียชีวิตจากอุบัติเหตุของประเทศไทย ในปี พ.ศ. 2537 ถึง 2545 ซึ่งได้จากการทดลองปัวซง (ชูเกียรติ ผุดพรมราช, 2547) การทดลองแบบปัวซง มีลักษณะดังนี้

1.1) สามารถแบ่งช่วงเวลาหรืออาณาบริเวณที่สนใจออกเป็นส่วนย่อยเล็กๆ ได้เป็นจำนวนมาก จนทำให้ความน่าจะเป็นของการเกิดเหตุการณ์ 1 ครั้ง ในแต่ละส่วนย่อยเล็กๆ มีค่าน้อยมาก

1.2) ความน่าจะเป็นของการเกิดเหตุการณ์ 1 ครั้ง ในแต่ละส่วนย่อยเล็กๆ จะเป็นสัดส่วนกับความยาวของช่วงเวลา หรือขนาดของอาณาบริเวณที่สนใจ ทั้งนี้จำนวนครั้งของเหตุการณ์ที่เกิดขึ้นช่วงนอกเวลา หรืออาณาบริเวณที่สนใจไม่มีผลกระทบต่อความน่าจะเป็นดังกล่าว

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

1.3) ความน่าจะเป็นของการเกิดเหตุการณ์มากกว่า 1 ครั้ง ในแต่ละส่วนย่อยเล็กๆ มีค่าน้อยมากจนละทิ้งได้ ซึ่งหมายความว่าในส่วนย่อยเล็กๆ นั้น จำนวนครั้งของเหตุการณ์ที่เกิดขึ้นอาจเป็น 0 หรือ 1 เท่านั้น

1.4) ในแต่ละส่วนย่อยเล็กๆ จำนวนครั้งของเกตุการณ์ที่เกิดขึ้นเป็นอิสระต่อกัน

ถ้า y เป็นจำนวนครั้งของเหตุการณ์ที่เกิดขึ้นในช่วงเวลาหรืออาณาบริเวณที่สนใจที่ได้จากการทดลองปัวซง Y จะเป็นตัวแปรสุ่มปัวซง และเรียกการแจกแจงความน่าจะเป็นของ X ซึ่งมีฟังก์ชันความน่าจะเป็น $f(y,\lambda)$ ดังนี้

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad ; y = 0,1,2,...$$

โดยที่ค่าคาดหมาย $E(Y) = \lambda$

และความแปรปรวน $V(Y) = \lambda \qquad ; \lambda > 0$

2.1.2 ตัวแบบการถดถอยที่เกี่ยวข้อง

1. ตัวแบบการถดถอยปัวซง

ตัวแบบการถดถอยปัวซงเป็นการหาความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระ โดยตัวแปรตามเป็นตัวแปรไม่ต่อเนื่องที่มีการแจกแจงปัวซง ส่วนตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพหรือตัวแปรเชิงปริมาณมีข้อตกลงเบื้องต้นดังนี้

$$E(Y) = V(Y)$$

ตัวแบบการถดถอยปัวซง

$$\underset{\sim}{\mathbf{y}} = e^{\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\underset{\sim}{\boldsymbol{\varepsilon}}}$$

โดยที่ $E(\mathbf{y}) = \boldsymbol{\lambda}$

จะได้ว่า $\underset{\sim}{\boldsymbol{\lambda}} = e^{\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\underset{\sim}{\boldsymbol{\varepsilon}}}$ เป็นตัวแบบการถดถอยปัวซง

และ $\underset{\sim}{\hat{\boldsymbol{\lambda}}} = e^{\underset{\sim}{\mathbf{X}}\underset{\sim}{\mathbf{b}}}$ เป็นสมการพยากรณ์ของการถดถอยปัวซง

จากตัวแบบการถดถอยปัวซงและสมการพยากรณ์ของการถดถอยปัวซง สามารถจัดอยู่ในรูปเชิงเส้นได้ดังนี้

$\ln \underset{\sim}{\boldsymbol{\lambda}} = \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\underset{\sim}{\boldsymbol{\varepsilon}}$ ซึ่งเป็นตัวแบบการถดถอยปัวซงในรูปเชิงเส้น

และ $\ln \underset{\sim}{\hat{\boldsymbol{\lambda}}} = \underset{\sim}{\mathbf{X}}\underset{\sim}{\mathbf{b}}$ เป็นสมการพยากรณ์ของการถดถอยปัวซงในรูปเชิงเส้น

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

เมื่อกำหนดให้ $\underset{\sim}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$ ขนาด n × 1

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$ ขนาด n × (k+1)

$\underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$ ขนาด (k+1) × 1 , $\underset{\sim}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix}$ ขนาด (k+1) × 1

$\underset{\sim}{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ . \\ \lambda_n \end{bmatrix}$ ขนาด n × 1 $\underset{\sim}{\hat{\lambda}} = \begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ . \\ . \\ . \\ \hat{\lambda}_n \end{bmatrix}$ ขนาด n × 1

และ $\underset{\sim}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$ ขนาด n × 1

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

โดยที่ $\underset{\sim}{\mathbf{y}}$ เป็นเวกเตอร์ แถวตั้งขนาด n ที่มีสมาชิกที่ i เป็น y_i เมื่อ $i = 1,2,...,n$

$\mathbf{X}$ เป็นเมทริกซ์ ขนาด n แถว และ k+1 หลัก

$\underset{\sim}{\boldsymbol{\beta}}$ เป็นเวกเตอร์ แถวตั้งขนาด k+1 ที่มีสมาชิกที่ i เป็น β_i

$\underset{\sim}{\mathbf{b}}$ เป็นเวกเตอร์ แถวตั้งขนาด k+1 ที่มีสมาชิกที่ i เป็น b_i

$\underset{\sim}{\lambda}$ เป็นเวกเตอร์ แถวตั้งขนาด n ที่มีสมาชิกที่ i เป็น λ_i

$\underset{\sim}{\hat{\lambda}}$ เป็นเวกเตอร์ แถวตั้งขนาด n ที่มีสมาชิกที่ i เป็น $\hat{\lambda}_i$

และ $\underset{\sim}{\boldsymbol{\varepsilon}}$ เป็นเวกเตอร์ แถวตั้งขนาด n ที่มีสมาชิกที่ i เป็น ε_i

2.1.3 ข้อมูลสูญหาย (Missing Data)

ข้อมูลสูญหาย คือ ค่าสังเกตที่ต้องการทราบค่าแต่ไม่สามารถทราบค่าได้ ข้อมูลสูญหายสามารถส่งผลกระทบต่อการวิจัยทั้งในส่วนของการวิเคราะห์และการสรุปผลตีความ โดยที่ระดับความรุนแรงของผลกระทบนี้ขึ้นอยู่กับองค์ประกอบจากหลายส่วนแต่ที่สำคัญก็คือ ขนาดของข้อมูลสูญหาย ประเภทของข้อมูลสูญหายที่เกิดขึ้นและวิธีการจัดการกับข้อมูลสูญหาย

2.1.4 ประเภทของข้อมูลสูญหาย (Type of Missing Data)

การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญ ทั้งนี้เพราะหากสามารถทราบถึงลักษณะชองข้อมูลสูญหายจะช่วยในการพิจารณาแนวทางสำหรับจัดการกับปัญหาความไม่สมบูรณ์ของข้อมูลได้อย่างเหมาะสม ซึ่งโดยทั่วไปมักจำแนกข้อมูลสูญหายออกเป็น 3 ประเภทด้วยกัน (ปิยะภรณ์ ประสิทธิ์วัฒนเสรีและสุคนธ์ ประสิทธิ์วัฒนเสรี, 2552) คือ

1) Missing Completely at Random (MCAR) เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด นั่นคือที่สูญหายเป็นอิสระจากตัวแปรต่างๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลปกติและข้อมูลสูญหายในกรณีนี้เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่างๆ ในฐานข้อมูล

2) Missing at Random (MAR) เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือค่าของข้อมูลสูญหายขึ้นอยู่กับตัวแปรอื่นๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย ยกตัวอย่างเช่น หากพบว่าเฉพาะกลุ่มผู้ที่ได้รับการศึกษาน้อยที่ไม่ได้ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการเสพยาเสพติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพติดมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับการศึกษา

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3) Not Missing at Random (NMAR) เป็นลักษณะของข้อมูลซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรตัวอื่นด้วย เช่น หากข้อมูลสูญหายของระดับรายได้ขึ้นอยู่กับระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นจัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใดๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวเริ่มต้นแต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรนอกฐานข้อมูล

2.1.5 การประมาณค่าสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation)

การศึกษานี้เป็นการประมาณค่าสูญหายภายใต้การถดถอยปัวซง โดยกำหนดให้มีขนาดตัวอย่าง n สุ่มจากประชากรทั้งหมด N หน่วย ประกอบด้วยตัวแปรอิสระ X สามตัวแปร มีขนาดสมบูรณ์จำนวน n หน่วย (x_{ki}) ; $k=1,2,3$ และ $i=1,...,r$ ข้อมูลตัวแปรตาม Y มีขนาดสมบูรณ์จำนวน r หน่วย และมีหน่วยตัวอย่าง $n-r$ หน่วยที่ไม่สมบูรณ์ นั่นคือ $y_{i'}$ เป็นข้อมูลสูญหาย ซึ่งแสดงได้ดังนี้

ตารางที่ 2.1 แสดงตัวแปรอิสระและตัวแปรตามแบบสมบูณณ์และแบบที่มีการสูญหาย

ตัวแปรตาม	y_1	y_2	. . .	y_r	y_{r+1}
ตัวแปรอิสระ	$x_{1.1}$	$x_{1.2}$	. . .	$x_{1.r}$	$x_{1.r+1}$
	$x_{2.1}$	$x_{2.2}$	. . .	$x_{2.r}$	$x_{2.r+1}$
	$x_{3.1}$	$x_{3.2}$	. . .	$x_{3.r}$	$x_{3.r+1}$

กำหนด (x_{ki}, y_i) แทนค่าตัวแปรอิสระ x_i ที่สมบูรณ์และมีตัวแปรตาม y_i ที่สมบูรณ์ ; $i=1,...,r$

$(x_{ki'}, y_{i'})$ แทนค่าตัวแปรอิสระ $x_{i'}$ ที่สมบูรณ์และมีตัวแปรตาม $y_{i'}$ ที่สูญหาย ; $i'=r+1,...,n$

การศึกษานี้มีวิธีการประมาณค่าสูญหายทั้งหมด 3 วิธี ได้แก่ 1. วิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation 2. วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation 3. วิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson Regression Imputation

1. วิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation (NNI)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Nearest Neighbor Imputation เป็นวิธีการหนึ่งที่จัดอยู่ในวิธี Hot Deck (Hot Deck Method) คือ เป็นการแทนข้อมูลที่สูญหายด้วยข้อมูลที่ทราบค่าโดยวิธีนี้เป็นวิธีที่มีประสิทธิภาพมากกว่าวิธี Hot Deck อื่นๆ (เช่น Sequence Hot - decking , Random Hot – decking) ซึ่งวิธี Nearest Neighbor Imputation จะประมาณค่าโดยใช้ค่าใกล้สุดโดยพิจารณาเลือกหน่วยตัวอย่างจากชุดข้อมูล $(x_{1i}, x_{2i} ... x_{ki}, y_i)$ ที่ทราบค่าที่มีลักษณะคล้ายคลึงกับหน่วยตัวอย่างที่เกิดค่าสูญหายมากที่สุด จากนั้นแทนค่าข้อมูลสูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงกัน (เพียงออ ยีสา, 2551)

กำหนดให้ $y_{i'}^{NNI} \quad ; i' = r+1, .., n$ คือค่าที่ได้จากการประมาณแบบ NNI พิจารณาความคล้ายจากระยะทางยุคลิด (Euclidean Distance) ดังนี้

กำหนดให้ $dist(x_{ki}, x_{ki'}) = \sqrt{\sum_{k=1}^{3} (x_{ki} - x_{ki'})^2}$ สำหรับ $i, 1 \le i \le r$ และ $i' = r+1, ..., n$

เมื่อ $y_i = \min_{1 \le i \le r} (dist(x_{ki}, x_{ki'}))$

จะได้ว่า $y_{i'}^{NNI} = y_i$

2. วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI)

การประมาณค่าสูญหายแบบ Poisson Regression Imputation (PRI) จัดเป็นเทคนิคที่ใช้ในการประมาณค่าสูญหายที่มีแนวคิดการใช้แบบ Regression Imputation ซึ่งเป็นเทคนิคพื้นฐานโดยใช้การถดถอยปัวซงจากชุดข้อมูล $(x_{1i} x_{2i}, ..., x_{ki}, y_i) \quad ; i = 1, ..., r$ ที่ทราบค่า เมื่อตัวแปรตามส่วนหนึ่งมีค่าสูญหาย $(y_{i'}^{PRI}) \quad ; i' = r+1, ..., n$ จะพยากรณ์ค่าสูญหายด้วยสมการถดถอยปัวซงโดยคำนวณค่าสัมประสิทธิ์การถดถอยจากชุดข้อมูล $(x_{1i} x_{2i}, ..., x_{ki}, y_i)$ ที่เหลืออยู่ นั่นคือ

ส่วนที่1 สมการการถดถอยปัวซงสำหรับข้อมูลที่สมบูรณ์

$$\hat{\lambda}_i = e^{b_0 + b_1 x_{1i} + ... + b_k x_{ki}} \quad ; i = 1, ..., r$$

ส่วนที่2 การพยากรณ์ค่าสูญหายด้วยสมการถดถอยปัวซง

$$y_{i'}^{PRI} = e^{b_0 + b_1 x_{1i'} + ... + b_k x_{ki'}} \quad ; i' = r+1, ..., n$$

3. วิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson Regression Imputation (WNPR)

วิธี Weighted Nearest Neighbor and Poisson Regression Imputation เป็นวิธีการประมาณค่าสูญหายที่พัฒนาโดยนำแนวคิดของสองวิธีการประมาณค่าสูญหายด้วยการรวมเข้าด้วยกันโดยรวม

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

วิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation และวิธี Poisson Regression Imputation แล้วถ่วงน้ำหนัก

กำหนดให้ $y_{i'}^{WNPR} ; i' = r+1,...,n$ คือค่าที่ได้จากวิธีการประมาณค่าสูญหายแบบ (WNPR)

$$y_{i'}^{WNPR} = w_{i'} y_{i'}^{NNI} + (1-w_{i'}) y_{i'}^{PRI}$$

เมื่อ $W_{i'} \sim U(0,1)$

$$w_{i'} = \begin{cases} 0 & U \le 0.5 \\ 1 & U > 0.5 \end{cases}$$

โดยที่ $w_{i'} ; i' = r+1,...,n$ เป็นค่าที่ใช้ในการถ่วงน้ำหนักระหว่าง NNI และ PRI

2.1.6 เกณฑ์ในการตัดสินใจ

เกณฑ์ในการการตัดสินใจว่าการประมาณค่าสูญหายว่าวิธีใดประมาณค่าได้ดีกว่า จากการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงในรูปแบบ Mean Square Error (MSE) โดยสามารถคำนวณจากสมการดังนี้

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

เมื่อ y_i แทนค่าจริงของข้อมูลตัวแปรตามที่ i

$\hat{y}_i$ แทนค่าประมาณของข้อมูลตัวแปรตามตัวที่ i

n แทนขนาดของตัวอย่าง

ในการเปรียบเทียบวิธีการประมาณ จะใช้ค่าเฉลี่ยของแต่ละ MSE ที่ได้ในแต่ละรอบมาเปรียบเทียบ ถ้าวิธีการประมาณใดให้ค่าเฉลี่ย MSE ต่ำกว่า แสดงว่าเป็นวิธีการประมาณที่ดีกว่า

2.2 งานวิจัยที่เกี่ยวข้อง

Fatma El Zahraa S. Salama, Ahmed M. Gad และ M. Mohamed (2016) ได้ทำการศึกษาเปรียบเทียบแบบจำลองข้อมูลระยะยาวแบบนับที่มีค่าสูญหาย โดยเปรียบเทียบภายใต้การแจกแจงแบบปัวซง มีวัตถุประสงค์เพื่อเปรียบเทียบ วิธีการวิเคราะห์ที่แตกต่างกันสำหรับข้อมูลจำนวนนับแบบระยะยาวที่มีข้อมูลสูญหาย โดยศึกษาจำลองเพื่อเปรียบเทียบวิธีการต่างๆ ภายใต้สถานการณ์ที่แตกต่างกัน นอกจากนี้ยังศึกษาในข้อมูลจริงด้วย กำหนดวิธีการสูญหายของข้อมูลด้วยวิธี MCAR และ วิธี MAR ขนาดตัวอย่างเป็น 20 50 และ 100 อัตราการสูญหายเป็นต่ำและสูง วิธีการสูญหายคือ วิธีการประมาณสมการนัยทั่วไป (GEE) วิธีการถ่วงน้ำหนักการประมาณสมการนัยทั่วไป (WGEE) และ วิธีการประมาณค่าจากการคำนวณหลายครั้งด้วยการผูกสมการ (MICE) เปรียบเทียบวิธีประมาณโดย วิธีเอนเอียงสัมพันธ์ (RB) ผลลัพธ์ในตารางที่ 2

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

จากผลลัพธ์ในกลุ่ม MCAR จะเห็นได้ว่าวิธีการทั้งหมดสร้างการประมาณที่ค่อนข้างเป็นกลาง ความเอนเอียงสัมพัทธ์ต่ำกว่า 30 % สำหรับวิธีการทั้งหมดที่มีการเปลี่ยนแปลงเล็กน้อย เมื่ออัตราการสูญหายและขนาดตัวอย่างที่เปลี่ยนแปลงไป ภายใต้ MCAR วิธี WGEE จะเอนเอียงน้อยกว่า และภายใต้ MAR วิธี WGEE นั้นดีกว่าในหลายกรณี

พัชรีพร สนรักษา, พัชรี วงษ์เกษม และ คณินทร์ ธีรภาพโอฬาร (2555) ได้ศึกษาการประมาณและการเปรียบเทียบตัวประมาณค่าเฉลี่ยในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหายในการวิเคราะห์สมการถดถอยเชิงเส้นอย่างง่าย คือ วิธีประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน วิธีประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ วิธีประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก วิธีการหาตัวประมาณค่าของ Singh และวิธีการปรับตัวประมาณของค่าใหม่ โดยใช้เกณฑ์การเปรียบเทียบประสิทธิภาพโดยใช้ค่า The root mean square error (RMSE) ผลการทดลองพบว่า วิธีการปรับตัวประมาณของค่าใหม่มีประสิทธิภาพดีกว่าวิธีการประมาณข้อมูลสูญหายวิธีอื่นๆ

ภาคย์ สิทธิผกาผล (2555) ได้ศึกษาการเปรียบเทียบวิธีการประมาค่าที่เก็บไม่ได้ 3 วิธี ได้แก่ วิธีการถดถอยพหุเชิงเส้น วิธีระยะห่างสุด และวิธีค่าเฉลี่ย โดยการจำลองประชากร ขนาด 5,000 หน่วย แล้วทำการสุ่มตัวอย่างขนาด 60, 80, 100, 150, 300 และ 500 มีตัวแปรที่ศึกษา Y และตัวแปรอิสระ X 6 ตัว สัมประสิทธิ์สหสัมพันธ์ระหว่าง X กับ Y อยู่ 3 ระดับคือระดับต่ำ ระดับปานกลางและระดับสูง โดยใช้โปรแกรมอาร์ สุ่มตัวอย่างซ้ำ 1,000 ครั้งจากแต่ละประชากรให้มีขนาดเล็ก ขนาดกลางและขนาดใหญ่สุ่มค่าสังเกต Y ในตัวอย่างทิ้ง 5, 10, 15 และ 20 เปอร์เซ็นต์ถือว่าเป็นค่าที่เก็บรวบรวมไม่ได้ ผลการศึกษาพบว่าค่าสัมประสิทธิ์การแปรผันของตัวประมาณมีแนวโน้มลดลงเมื่อมีขนาดตัวอย่างและร้อยละของข้อมูลที่สูญหายเพิ่มขึ้น วิธีการถดถอยพหุเชิงเส้นให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณต่ำสุดและวิธีค่าเฉลี่ย ให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณสูงสุด เมื่อตัวแปร Y และตัวแปร X 6 ตัวมีสัมประสิทธิ์สหสัมพันธ์ในระดับปานกลางและระดับสูง วิธีการถดถอยพหุเชิงเส้นให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณต่ำสุดและวิธีการประมาณแบบวิธีระยะห่างต่ำสุดให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณสูงสุด เมื่อตัวแปร Y และตัวแปร X 6 ตัว มีสัมประสิทธิ์สหสัมพันธ์ในระดับต่ำ

อุษณีย์ วงศ์อามาตย์ (2555) ได้ทำการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงเส้นพหุ วิธีการประมาณค่าสูญหายที่ใช้ในงานการวิจัยนี้คือ วิธี EM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองโดยมีสัดส่วนของการสูญหาย 3 ระดับ คือ 10%, 20%, 30% และมีระดับของการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับคือ ไม่มี, ปานกลาง,

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

สูง จากการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (average mean square error; AMSE) พบว่า i) วิธีการประมาณทุกวิธีสามารถประมาณได้ดีขึ้นเมื่อขนาดตัวอย่างมีขนาดใหญ่ขึ้น ii) วิธีการประมาณทุกวิธีประมาณได้แย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน สัดส่วนของการสูญหาย และ ระดับของการสูญหายแบบอิกนอร์เรเบิล มีค่าเพิ่มขึ้น iii) โดยรวมแล้วิธี EM ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าไม่สูง (10-30) และ iv) วิธี KNN ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าสูง (90)

ทัตดา หิรัญพต (2554) ได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย (Imputation Method) ระหว่าง วิธีการประมาณค่าข้อมูลสูญหาย ด้วยค่าถดถอย (RI) กับวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก (SRI) โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 100, 200 และ 300 สุ่มโดยขั้นตอนวิธีของเบ็บบิงตัน (Bebbington algotithm) ซึ่งใช้วิธีการชักตัวอย่างแบบสุ่มเชิงเดียวชิดไม่คืนที่ (Simple random sampling without replacement หรือ SRSWOR) และกำหนดเปอร์เซ็นต์ของข้อมูลสูญหายของตัวอย่าง เท่ากับ 5%, 10%, 15% และ 20% ของขนาดตัวอย่าง ผลการศึกษาการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณค่าข้อมูลสูญหายนั้น ได้วัดจาก ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณของข้อมูลสูญหาย (MSE) โดยได้ข้อสรุปว่า วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย (RI) มีประสิทธิภาพมากกว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก (SRI)

เพียงออ ยีสา (2551) ได้ทำการเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นเพื่อการพยากรณ์โดยพิจารณาข้อมูล 2 ลักษณะ คือข้อมูลภาคตัดขวาง และข้อมูลอนุกรมเวลาที่มีปัจจัยด้านแนวโน้มและปัจจัยฤดูกาลเข้ามาเกี่ยวข้องโดยทำการประมาณค่าสูญหายด้วยวิธีการ Regression Imputation (RI) วิธี Nearest Neighbor Imputation (NNI) วิธี Weighted Nearest Neighbor and Regression Imputation (WNR) และวิธี EM algorithm (EM) โดยจำลองข้อมูลซึ่งกำหนดขนาดตัวอย่าง 50, 100 และ 200 ส่วนเบี่ยงเบนมาตรฐาน 5, 10, 15, 20 และ 25 ร้อยละการสูญหายของตัวแปรเป็น 5, 10 และ 20 ตามลำดับ ซึ่งกำหนดให้ตัวแปรอิสระมีการแจกแจงปกติ ซึ่งเกณฑ์ในการเปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายจะใช้ค่า MAPE จากการศึกษาได้ผลดังนี้ สำหรับข้อมูลภาคตัดขวาง กรณีที่ค่าสหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทั้ง 2 ตัวสูง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำถึงปานกลาง วิธี RI และ EM มีค่า MAPE ต่ำกว่าวิธีอื่น เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูงวิธี WNR ให้ผลดีกว่าวิธีอื่น กรณีที่ค่าสหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระตัวหนึ่งสูงมากและอีกตัวหนึ่งปานกลาง วิธี RI และ EM จะให้ผลดีกว่าวิธีอื่นๆ สำหรับข้อมูลอนุกรมเวลา วิธี WNR มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่นๆ ในกรณีที่ข้อมูลที่มีอิทธิพลของฤดูกาลสูง และวิธี NNI จะให้ผลดีเมื่อส่วนเบี่ยงเบน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

มาตรฐานอยู่ในระดับสูง สำหรับข้อมูลที่มีอิทธิพลจากปัจจัยแนวโน้มสูงวิธี RI และ EM เป็นวิธีที่ให้ผลดีกว่าวิธีอื่น กรณีที่ข้อมูลมีอิทธิพลจากปัจจัยแนวโน้มและปัจจัยฤดูกาลระดับปานกลาง เมื่อส่วนเบี่ยงเบนมาตรฐานเพิ่มสูงขึ้นวิธี WNR เป็นวิธีที่มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่นๆ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บทที่ 3

วิธีการดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงทดลอง มีวัตถุประสงค์ เพื่อเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายในการถดถอยปัวซง โดยทำการประมาณค่าสูญหายของตัวแปรตามสามวิธี ได้แก่ 1. วิธีการประมาณค่าสูญหายแบบระยะห่างใกล้สุด (Nearest Neighbor Imputation: NNI) 2. วิธีการประมาณค่าสูญหายแบบการถดถอยปัวซง (PoissonRegression Imputation: PRI) 3. วิธีการประมาณค่าสูญหายแบบถ่วงน้ำหนัก (Weighted Nearest Neighbor and PoissonRegression Imputation: WNPR) โดยทำการเปรียบเทียบด้วยวิธีค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริง ด้วยวิธีค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) ดังนั้นในบทนี้จะกล่าวถึงวิธีการประมาณค่าสูญหายและขั้นตอนการดำเนินการศึกษา

3.1 แผนการดำเนินการวิจัย

การวิจัยนี้เป็นการประมาณค่าสูญหายของตัวแปรตามในการถดถอยปัวซง เมื่อข้อมูลตัวแปรตามมีการสูญหายแบบสุ่ม (Missing At Random: MAR) โดยทำการประมาณค่าสูญหายของตัวแปรตามทั้งสามวิธี ที่ขนาดตัวอย่าง 4 ระดับ ค่าร้อยละการสูญหายของตัวแปรตามสามระดับ ค่าพารามิเตอร์ ประกอบไปด้วยความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับตัวแปรตามทั้งหมด 5 ชุด และตัวแปรอิสระจำนวนสามตัว จากนั้นเปรียบเทียบความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริง ในรูปแบบ MSE ของค่าประมาณที่ได้จากการประมาณค่าสูญหายทั้งสามวิธี เพื่อหาวิธีที่เหมาะสมในแต่ละสถานการณ์

3.2 ขั้นตอนการดำเนินการวิจัย

การวิจัยครั้งนี้จะสร้างข้อมูลตัวแปรตามและตัวแปรอิสระ ภายใต้การแจกแจงถดถอยปัวซง (Poisson Regression) โดยทำการสร้างแบบจำลองสถานการณ์ด้วยโปรแกรม R Version 3.3.1 ซึ่งดำเนินการดังนี้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

1. ประชากร ในการศึกษาครั้งนี้ผู้วิจัยได้กำหนดให้ข้อมูลตัวแปรตามที่มีการแจกแจงปัวซง
2. กลุ่มตัวอย่างที่ถูกจากประชาการที่มีการแจกแจงปัวซง เท่ากับ 20, 60, 100 และ 200 หน่วย
3. จำลองข้อมูล กำหนดค่าพารามิเตอร์ β ประกอบด้วยพารามิเตอร์จำนวน 5 ชุด ขนาดตัวอย่าง 20, 60, 100 และ 200 หน่วย ค่าร้อยละการสูญหายเป็น 5 10 และ 15

   3.1 พารามิเตอร์ชุดที่ 1 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$

   3.2 พารามิเตอร์ชุดที่ 2 คือ $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$

   3.3 พารามิเตอร์ชุดที่ 3 คือ $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$

   3.4 พารามิเตอร์ชุดที่ 4 คือ $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$

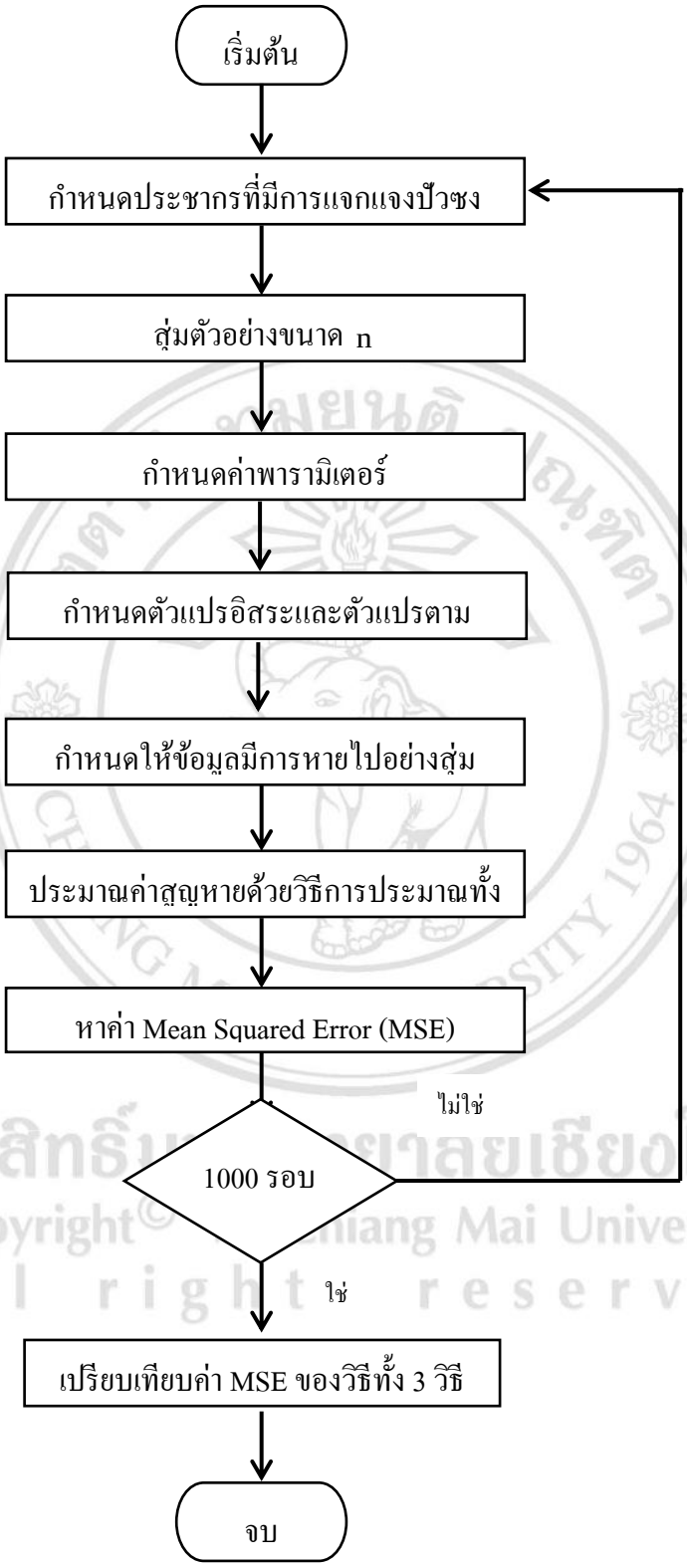
   3.5 พารามิเตอร์ชุดที่ 5 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$

4. กำหนดตัวแปรประกอบไปด้วยตัวแปรอิสระ X จำนวน 3 ตัวแปร โดยที่ ที่มีการแจกแจงเอกรูป U(0,1) โดยที่มีพารามิเตอร์ λ โดยที่ $\lambda = e^{x\beta}$ และตัวแปรตาม Y ที่มีการแจกแจงแบบปัวซง (n, λ)
5. กำหนดให้ข้อมูลมีการหายไปอย่างสุ่ม (MAR) ในตัวแปรตาม Y โดยกำหนดร้อยละการสูญหายเท่ากับ 5, 10 และ 15
6. ประมาณค่าสูญหายด้วยวิธีการประมาณทั้งสามวิธี โดยสร้างแบบจำลองสถานการณ์ในโปรแกรม R
7. กำหนดจำนวนรอบของการทำซ้ำจำนวน 1,000 รอบ
8. เปรียบเทียบความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริง ในรูปแบบ MSE

รายละเอียดดังภาพที่ 3.1

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

การจำลองข้อมูลตามที่ได้กล่าวไว้ สามารถเขียนสรุปเป็นแผนผังการดำเนินงานได้ดังนี้



ภาพที่ 3.1 แผนผังการดำเนินงาน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บทที่ 4

ผลการศึกษา

ในบทนี้จะกล่าวถึงผลการศึกษาของการประมาณค่าสูญหายด้วยวิธีการประมาณค่าสูญหายทั้ง 3 วิธี โดยทำการจำลองข้อมูลให้มีการแจกแจงปัวซง ที่มีขนาดตัวอย่าง 20, 60, 100 และ 200 หน่วย และจำลองให้ข้อมูลมีการสูญหายไปอย่างสุ่มด้วยวิธี MAR โดยมีร้อยละของการสูญหายเป็น 5, 10 และ 15 ของขนาดตัวอย่าง ที่มีค่าพารามิเตอร์ดังต่อไปนี้

ชุดที่ 1 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$

ชุดที่ 2 คือ $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$

ชุดที่ 3 คือ $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$

ชุดที่ 4 คือ $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$

และชุดที่ 5 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$

มีการทำซ้ำจำนวน 1,000 รอบ โดยในแต่ละสถานการณ์ใช้แบบจำลองมอนติคาร์โล (Monte carlo Simulation Technique) พร้อมทั้งมีการเปรียบเทียบวิธีการประมาณค่าสูญหายของข้อมูลทั้ง 3 วิธี ด้วยวิธี Mean Squared Error (MSE)

ส่วนที่ 1 ผลการประมาณค่าสูญหายของข้อมูลด้วยวิธีการประมาณทั้งสามวิธี ในแต่ละสถานการณ์

ส่วนที่ 2 กราฟแสดงผลการประมาณค่าสูญหายทั้งสามวิธี

กำหนดสัญลักษณ์ที่ใช้ในการนำเสนอผลการศึกษาดังนี้

n	หมายถึง	ขนาดตัวอย่าง
β	หมายถึง	พารามิเตอร์แสดงขนาดและทิศทาง
NNI	หมายถึง	วิธีการประมาณค่าสูญหายใกล้สุด
PRI	หมายถึง	วิธีประมาณค่าสูญหายด้วยวิธีการถดถอยแบบปัวซง
WNPR	หมายถึง	วิธีการประมาณค่าสูญหายถ่วงน้ำหนัก
MSE	หมายถึง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

4.1 ผลการประมาณค่าสูญหายของข้อมูลด้วยวิธีการประมาณค่าทั้งสามวิธี ในแต่ละสถานการณ์

การประมาณค่าสูญหายของข้อมูลด้วยวิธีการประมาณค่าทั้งสามวิธีได้แก่ วิธีการประมาณค่าสูญหายแบบ Nearest Neighbor Imputation (NNI) วิธีการประมาณค่าสูญหายแบบ Poisson Regression Imputation และวิธีการประมาณค่าสูญหายแบบ Weighted Nearest Neighbor and Poisson Regression Imputation (WNPR) ในแต่ละสถานการณ์ตามค่าพารามิเตอร์ที่กำหนด

จากตารางที่ 4.1-4.5 เป็นการเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีต่างๆ และทำการเปรียบเทียบโดยใช้เกณฑ์เปรียบเทียบด้วยวิธี MSE โดยที่ n มีขนาด 20, 60, 100 และ 200 และค่าร้อยละของการสูญหายเป็น 5, 10 และ 15 พบว่าพารามิเตอร์ชุดที่ 1 ชุดที่ 2 ชุดที่ 3 และชุดที่ 5 วิธี NNI ให้ค่า MSE ต่ำสุด แต่เนื่องจากค่า MSE ที่ต่ำลง เมื่อร้อยละการสูญหายเพิ่มมากขึ้น ฉะนั้นจึงไม่แนะนำให้ใช้ และวิธี PRI ให้ค่า MSE ที่ต่ำที่สุดในพารามิเตอร์ชุดที่ 4 วิธี PRI จึงเหมาะสมในกรณี พารามิเตอร์ชุดที่ 4

เมื่อพิจารณาที่ค่าร้อยละการสูญหายพบว่า เมื่อร้อยละการสูญหายเพิ่มมากขึ้น วิธี PRI และวิธี WNPR จะให้ค่า MSE ที่เพิ่มขึ้นตามไปด้วย ในขณะที่ชุดพารามิเตอร์ที่ 1, 2, 3 และ 5 วิธี NNI ให้ค่า MSE ต่ำลง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 4.1 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายสามวิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์เป็น $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$

n	ร้อยละการสูญหาย	Method		
		NNI	PRI	WNPR
20	5	0.5691*	0.5948	0.5857
	10	0.5612*	0.6157	0.5964
	15	0.5530*	0.6452	0.6122
60	5	0.5560*	0.5824	0.5676
	10	0.5378*	0.5926	0.5849
	15	0.5153*	0.5898	0.5908
100	5	0.5542*	0.5803	0.5711
	10	0.5355*	0.5848	0.5880
	15	0.5150*	0.5902	0.5982
200	5	0.5666*	0.5933	0.5955
	10	0.5444*	0.5989	0.6219
	15	0.5156*	0.5949	0.6253

หมายเหตุ * หมายถึง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีในแต่ละสถานการณ์มีค่าน้อยที่สุด

พิจารณาตารางที่ 4.1 พารามิเตอร์ชุดที่ 1 ประกอบด้วย $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$ พิจารณาที่ n มีขนาด 20, 60, 100 และ 200 เมื่อค่าร้อยละการสูญหายเป็น 5,10 และ 15 พบว่าวิธี NNI ให้ค่า MSE ต่ำสุด รองลงมาคือ วิธี WNPR และ วิธี PRI ตามลำดับ เมื่อพิจารณาที่วิธี NNI เมื่อค่าร้อยละการสูญหายเพิ่มมากขึ้น ค่า MSE มีค่าที่ต่ำลง และพิจารณาที่ขนาดตัวอย่างพบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นค่า MSE ขึ้นลงไม่สม่ำเสมอ ในขณะที่วิธี PRI และวิธี WNPR เมื่อค่าร้อยละการสูญหายเพิ่มมากขึ้น ค่า MSE มีค่าที่เพิ่มขึ้นตามไปด้วย แต่ขนาดตัวอย่างเพิ่มมากขึ้นค่า MSE ขึ้นลงไม่สม่ำเสมอ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 4.2 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายสามวิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์เป็น $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$

n	ร้อยละการสูญหาย	Method		
		NNI	PRI	WNPR
20	5	0.7971*	0.8613	0.8338
	10	0.8132*	0.9110	0.8663
	15	0.8291*	0.9037	1.0407
60	5	0.7701*	0.8019	0.7882
	10	0.7547*	0.8429	0.8341
	15	0.7045*	0.8532	0.8807
100	5	0.7705*	0.8044	0.8049
	10	0.7340*	0.8105	0.8948
	15	0.7009*	0.8196	0.9455
200	5	0.7743*	0.8089	0.8892
	10	0.7321*	0.8102	0.9296
	15	0.6964*	0.8108	0.9493

หมายเหตุ *หมายถึง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีในแต่ละสถานการณ์มีค่าน้อยที่สุด

พิจารณาตารางที่ 4.2 พารามิเตอร์ชุดที่ 2 ประกอบด้วย $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$ พิจารณาที่ n มีขนาด 20, 60, 100 และ 200 เมื่อค่าร้อยละการสูญหายเป็น 5, 10 และ 15 พบว่าวิธี NNI ให้ค่า MSE ต่ำสุด รองลงมาคือ วิธี WNPR ในขนาดตัวอย่างที่ 20 และ 60 และวิธี PRI ในขนาดตัวอย่างที่ 100 และ 200

ขนาดตัวอย่างที่ 20 เมื่อร้อยละการสูญหายมากขึ้นพบว่าค่า MSE มีค่าสูงขึ้นในทั้ง 3 วิธี พิจารณาที่ n=60, 100 และ 200 ในวิธี NNI เมื่อร้อยละการสูญหายเพิ่มขึ้นค่า MSE มีค่าต่ำลง ในขณะที่วิธี PRI และ WNPR เมื่อร้อยละการสูญหายเพิ่มขึ้น พบว่าค่า MSE มีค่าสูงขึ้น

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 4.3 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายสามวิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์เป็น $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$

n	ร้อยละการสูญหาย	Method		
		NNI	PRI	WNPR
20	5	1.2929*	1.3684	1.3231
	10	1.3394*	1.6526	1.5153
	15	1.3740*	1.6823	1.6118
60	5	1.2629*	1.3381	1.3010
	10	1.0970*	1.3383	1.3308
	15	1.0831*	1.3479	1.3385
100	5	1.1414*	1.2286	1.1904
	10	1.0959*	1.2346	1.2804
	15	1.0465*	1.2407	1.3470
200	5	1.1274*	1.2168	1.2221
	10	1.0742*	1.2245	1.3222
	15	1.0136*	1.2271	1.3358

หมายเหตุ *หมายถึง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีในแต่ละสถานการณ์มีค่าน้อยที่สุด

พิจารณาตารางที่ 4.3 พารามิเตอร์ชุดที่ 3 ประกอบด้วย $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$ พิจารณาที่ n มีขนาด 20, 60, 100 และ 200 เมื่อค่าร้อยละการสูญหายเป็น 5, 10 และ 15 พบว่าวิธี NNI ให้ค่า MSE ต่ำสุด รองลงมาคือ วิธี WNPR ในขนาดตัวอย่างที่ 20, 60 และ 100 และวิธี PRI ในขนาดตัวอย่างที่ 200

เมื่อพิจารณาที่ n = 20 เมื่อร้อยละการสูญหายมากขึ้นพบว่าค่า MSE มีค่าสูงขึ้นในทั้งสามวิธี พิจารณาที่ n=60, 100 และ 200 ในวิธี NNI เมื่อร้อยละการสูญหายเพิ่มขึ้นค่า MSE มีค่าต่ำลง ในขณะที่วิธี PRI และ WNPR เมื่อร้อยละการสูญหายเพิ่มขึ้น พบว่าค่า MSE มีค่าสูงขึ้น

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 4.4 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายสามวิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์เป็น $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$

n	ร้อยละการสูญหาย	Method		
		NNI	PRI	WNPR
20	5	4.7496	4.5293*	4.7078
	10	5.9849	5.1315*	5.2375
	15	5.6332	5.1978*	5.4752
60	5	5.6359	5.3325*	5.4624
	10	6.4026	5.7103*	6.0182
	15	8.2922	7.2287*	7.7146
100	5	5.1385	4.7827*	4.9406
	10	5.5178	4.8510*	5.1636
	15	6.2169	4.9898*	5.5093
200	5	4.6752	4.3460*	4.5160
	10	5.0756	4.4137*	4.7524
	15	5.6226	4.5001*	5.0527

หมายเหตุ * หมายถึง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีในแต่ละสถานการณ์มีค่าน้อยที่สุด

พิจารณาตารางที่ 4.4 พารามิเตอร์ชุดที่ 4 ประกอบด้วย $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$ พิจารณาที่ n มีขนาด 20, 60, 100 และ 200 เมื่อค่าร้อยละการสูญหายเป็น 5, 10 และ 15 พบว่าวิธี PRI ให้ค่า MSE ต่ำสุด รองลงมาคือ วิธี WNPR และวิธี NNI ตามลำดับ เมื่อพิจารณาในทุกวิธีพบว่าเมื่อร้อยละการสูญหายเพิ่มขึ้น ค่า MSE จะมีค่าสูงขึ้นตามไปด้วย และเมื่อขนาดตัวอย่างตั้งแต่ 60 หน่วยขึ้นไป ค่า MSE จะมีค่าลดลง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 4.5 ค่าประมาณข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายสามวิธี ขนาดตัวอย่าง ร้อยละการสูญหาย เมื่อค่าพารามิเตอร์เป็น $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$

n	ร้อยละการสูญหาย	Method		
		NNI	PRI	WNPR
20	5	0.2109*	0.2190	0.2154
	10	0.1984*	0.2144	0.2063
	15	0.1898*	0.2122	0.2015
60	5	0.2251*	0.2343	0.2298
	10	0.2119*	0.2358	0.2289
	15	0.2085*	0.2359	0.2284
100	5	0.2303*	0.2395	0.2360
	10	0.2224*	0.2423	0.2355
	15	0.2092*	0.2436	0.2308
200	5	0.2321*	0.2417	0.2385
	10	0.2231*	0.2427	0.2379
	15	0.2165*	0.2490	0.2450

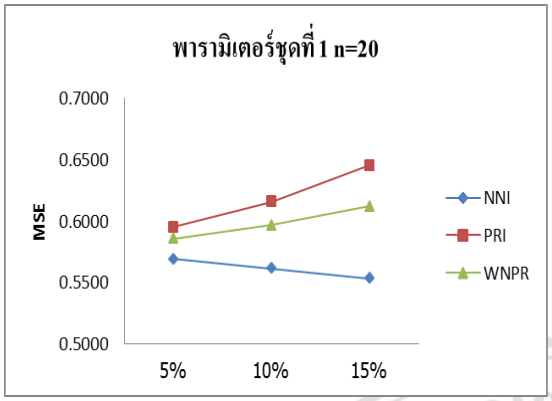
หมายเหตุ * หมายถึง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีในแต่ละสถานการณ์มีค่าน้อยที่สุด

พิจารณาตารางที่ 4.5 พารามิเตอร์ชุดที่ 5 ประกอบด้วย $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$ พิจารณาที่ n มีขนาด 60, 100 และ 200 เมื่อค่าร้อยละการสูญหายเป็น 5, 10 และ 15 พบว่าวิธี NNI ให้ค่า MSE ต่ำสุด รองลงมาคือ วิธี WNPR และวิธี PRI ตามลำดับ เมื่อร้อยละการสูญหายเพิ่มขึ้น วิธี NNI ให้ค่า MSE ที่ต่ำลงไปด้วย

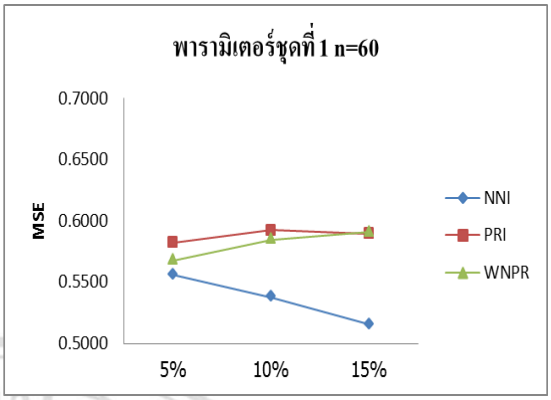
พิจารณาที่ n=20 เมื่อค่าร้อยละการสูญหายเพิ่มขึ้น PRI ให้ค่า MSE ต่ำลง และที่ขนาดตัวอย่างเป็น 60, 100 และ 200 จะให้ค่า MSE สูงขึ้น และเมื่อพิจารณาที่ 20, 60 และ 200 เมื่อค่าร้อยละการสูญหายเพิ่มขึ้น WNPR ให้ค่า MSE ต่ำลง และที่ขนาดตัวอย่างเป็น 100 จะให้ค่า MSE สูงขึ้น

เมื่อพิจารณาที่ขนาดตัวอย่าง ในทุกวิธี เมื่อขนาดตัวอย่างสูงขึ้นกลับให้ค่า MSE ที่สูงขึ้นตามไปด้วย

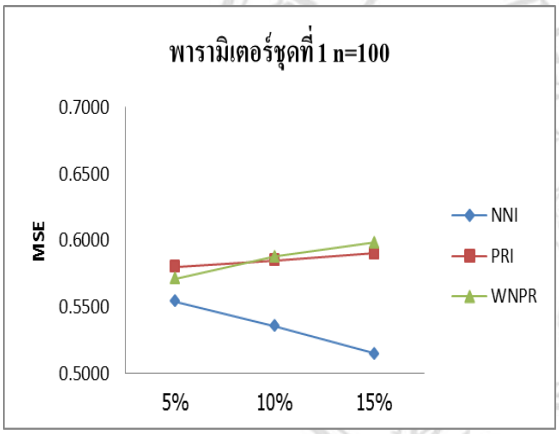
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



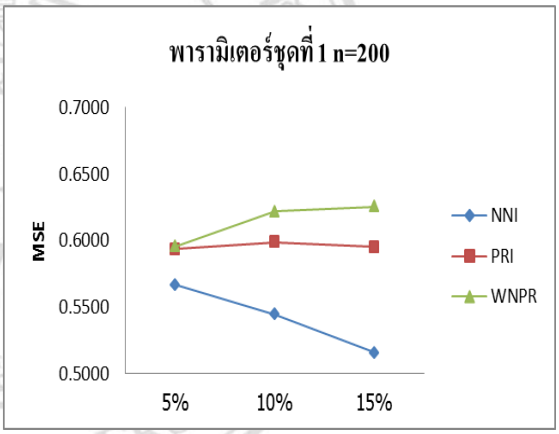
(ก)



(ข)



(ค)



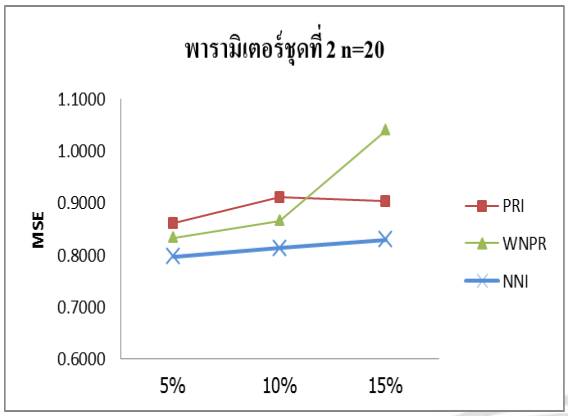
(ง)

ภาพที่ 4.1 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI, PRI และ WNPR ที่มีค่าพารามิเตอร์เป็น ($\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$)

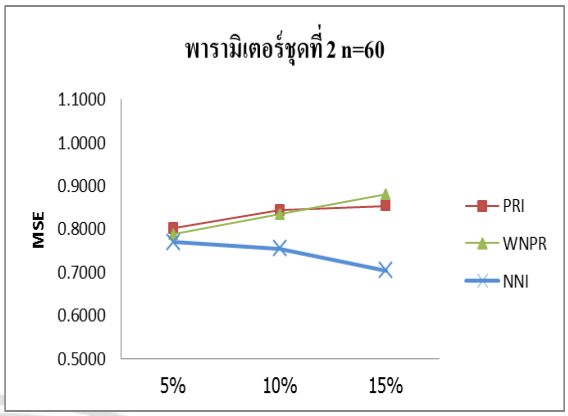
(ก) n = 20 (ข) n = 60

(ค) n = 100 (ง) n = 200

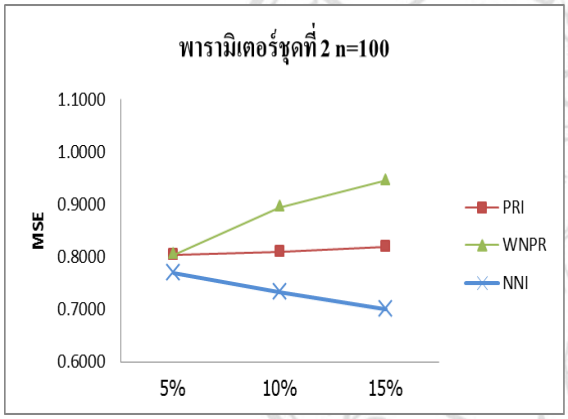
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



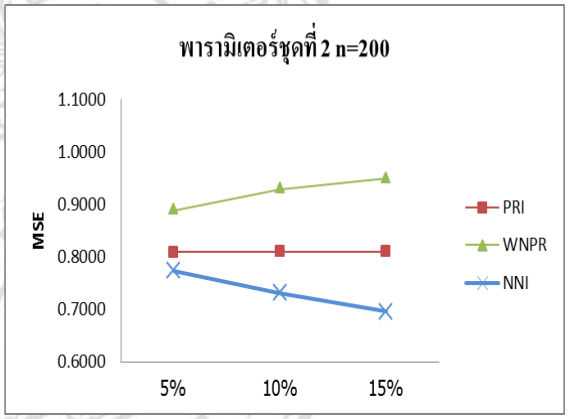
(ก)



(ข)



(ค)



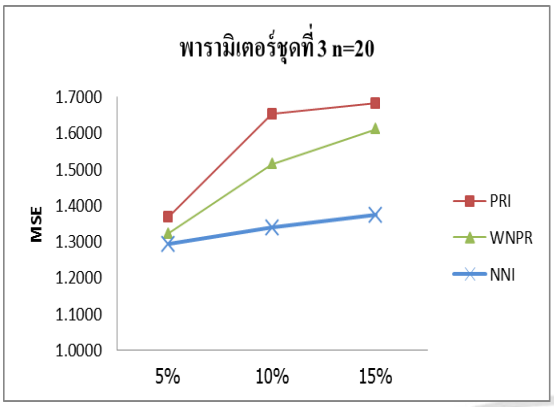
(ง)

ภาพที่ 4.2 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI, PRI และ WNPR ที่มีค่าพารามิเตอร์เป็น ($\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$)

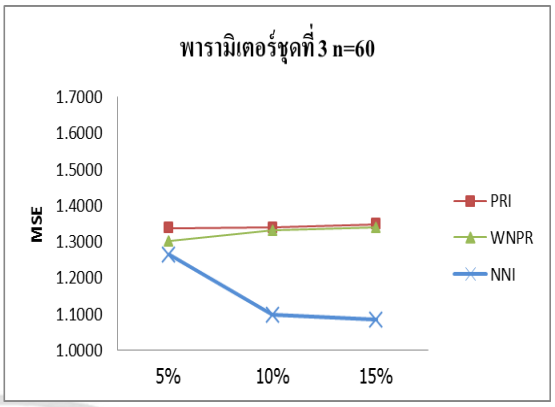
(ก) n = 20 (ข) n = 60

(ค) n = 100 (ง) n = 200

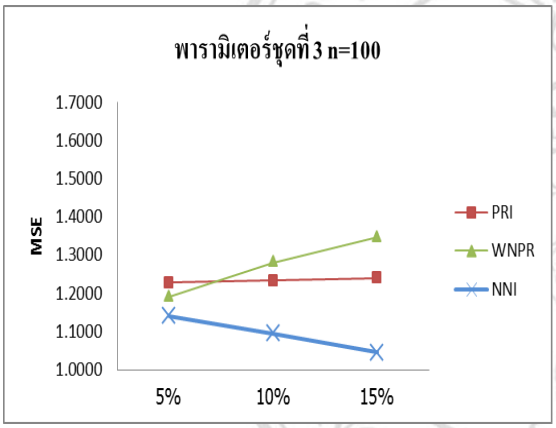
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



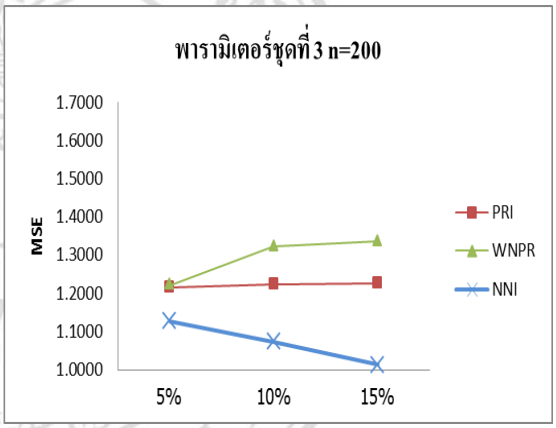
(ก)



(ข)



(ค)



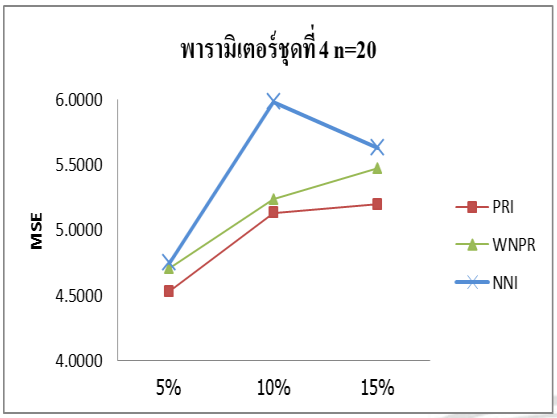
(ง)

ภาพที่ 4.3 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI, PRI และ WNPR ที่มีค่าพารามิเตอร์เป็น ($\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$)

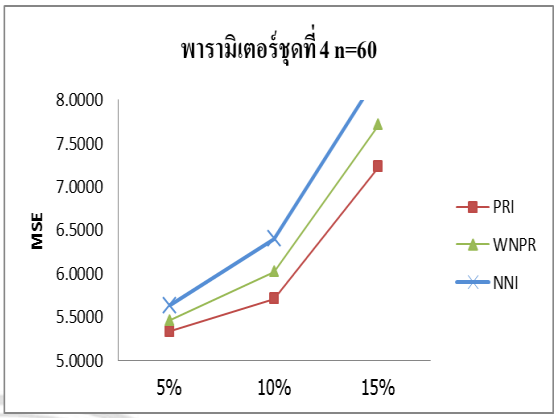
(ก) n = 20 (ข) n = 60

(ค) n = 100 (ง) n = 200

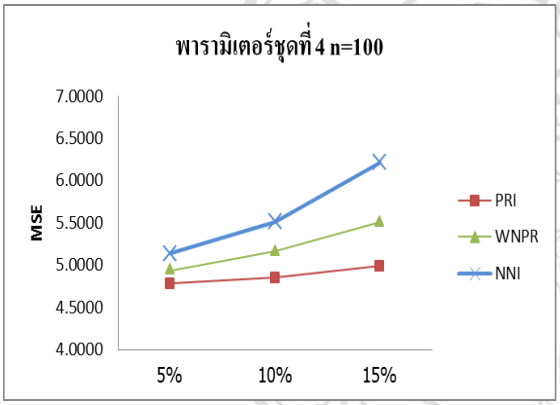
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



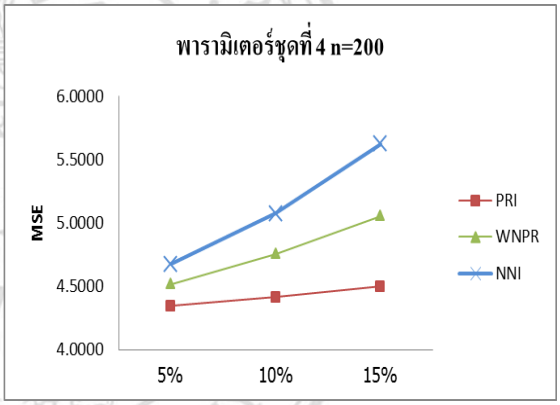
(ก)



(ข)



(ค)



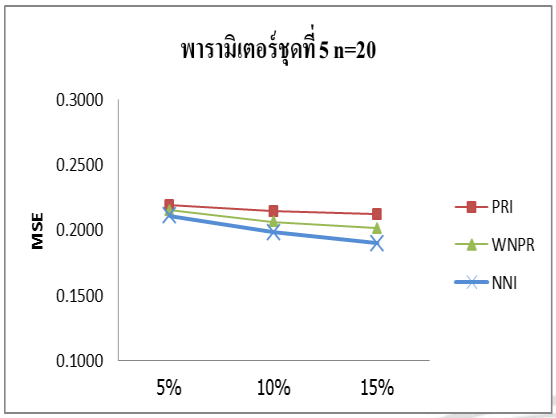
(ง)

ภาพที่ 4.4 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI, PRI และ WNPR ที่มีค่าพารามิเตอร์เป็น ($\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$)

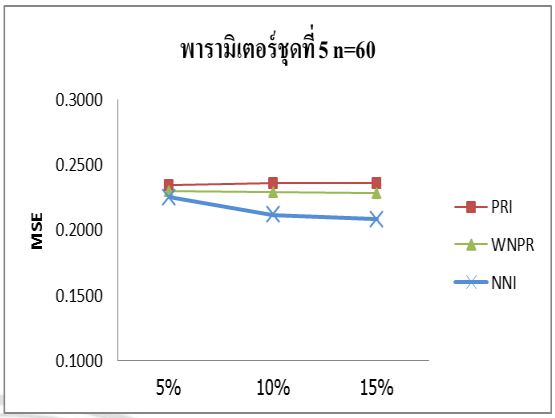
(ก) n = 20 (ข) n = 60

(ค) n = 100 (ง) n = 200

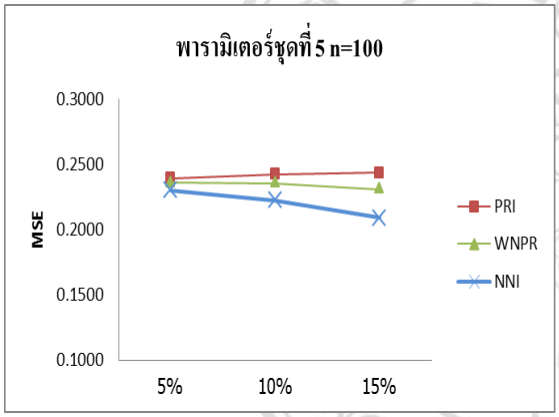
ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved



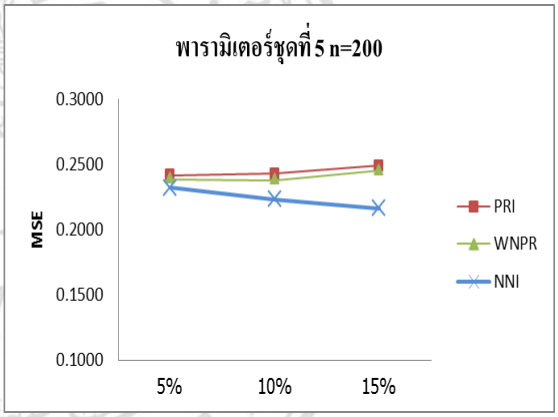
(ก)



(ข)



(ค)



(ง)

ภาพที่ 4.5 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการประมาณค่าสูญหาย NNI, PRI และ WNPR ที่มีค่าพารามิเตอร์เป็น ($\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$)

(ก) n = 20 (ข) n = 60

(ค) n = 100 (ง) n = 200

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

บทที่ 5

สรุปผลและอภิปราย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายในการวิเคราะห์การถดถอยป๊วซง โดยทำการจำลองสถานการณ์ให้ข้อมูลมีการสูญภายใต้การแจกแจงป๊วซงในกรณีขนาดตัวอย่าง และร้อยละการสูญหายต่างกัน พร้อมทั้งประมาณค่าสูญหายด้วยวิธีระยะห่างใกล้สุด วิธีการถดถอยป๊วซง และวิธีถ่วงน้ำหนัก จากนั้นทำการเปรียบเทียบวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยวิธีคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดซึ่งสามารถสรุปผลและอภิปรายผลการศึกษาได้ดังนี้

5.1 สรุปผลการศึกษา

5.1.1 ผลการศึกษาในการเปรียบเทียบค่าคลาดเคลื่อนกำลังสองเฉลี่ย

เมื่อจำลองข้อมูลภายใต้การแจกแจงป๊วซงมีขนาดตัวอย่าง 20, 60, 100 และ 200 ค่าร้อยละการสูญหายเป็น 5, 10 และ 15และมีพารามิเตอร์ดังต่อไปนี้

ชุดที่ 1 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = -0.2$ และ $\beta_3 = 0.5$
ชุดที่2 คือ $\beta_0 = 0.5, \beta_1 = -0.2, \beta_2 = 0.2$ และ $\beta_3 = 0.7$
ชุดที่ 3 คือ $\beta_0 = 0.2, \beta_1 = 0.5, \beta_2 = 07$ และ $\beta_3 = -0.2$
ชุดที่ 4 คือ $\beta_0 = -0.5, \beta_1 = -0.8, \beta_2 = -0.2$ และ $\beta_3 = -0.7$
และชุดที่ 5 คือ $\beta_0 = 0.7, \beta_1 = 0.2, \beta_2 = 0.8$ และ $\beta_3 = 0.5$

พารามิเตอร์ชุดที่ 1 พบว่าวิธีระยะห่างใกล้สุดเป็นวิธีที่ให้ค่าความเคลื่อนต่ำสุด รองลงมาคือวิธีถ่วงน้ำหนักและวิธีการถดถอยป๊วซงตามลำดับ เมื่อค่าร้อยละการสูญหายเพิ่มมากขึ้น วิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำลง และเมื่อขนาดตัวอย่างเพิ่มขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยขึ้นลงไม่สม่ำเสมอ ในขณะที่วิธีการถดถอยป๊วซงและวิธีถ่วงน้ำหนักเมื่อค่าร้อยละการสูญหายเพิ่มมากขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าที่เพิ่มขึ้นตามไปด้วย แต่ขนาดตัวอย่างเพิ่ม

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

มากขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยขึ้นลงไม่สม่ำเสมอ ในกรณีนี้วิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุด

พารามิเตอร์ชุดที่ 2 พบว่าวิธีระยะห่างใกล้สุดเป็นวิธีที่ให้ค่าความเคลื่อนกำลังสองเฉลี่ยต่ำสุดในขนาดตัวอย่างเป็น 20 เมื่อร้อยละการสูญหายมากขึ้นพบว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าสูงขึ้นในทั้ง 3 วิธี ส่วนขนาดตัวอย่างเป็น 60, 100 และ 200 ในวิธีระยะห่างใกล้สุดเมื่อร้อยละการสูญหายเพิ่มขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำลง ในขณะที่วิธีการถดถอยป๊วซงและวิธีถ่วงน้ำหนักเมื่อร้อยละการสูญหายเพิ่มขึ้น พบว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าสูงขึ้น ในกรณีวิธีการถดถอยป๊วซงจึงเป็นวิธีที่เหมาะสมที่สุดที่เมื่อขนาดตัวอย่างเป็น 100 และ 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 20 และ 60

พารามิเตอร์ชุดที่ 3 ขนาดตัวอย่างเป็น 20, 60 และ 100 พบว่าวิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ส่วนขนาดตัวอย่างเป็น 200 พบว่าวิธีการถดถอยป๊วซงให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด เมื่อพิจารณาที่ขนาดตัวอย่างเป็น 20 เมื่อร้อยละการสูญหายมากขึ้นพบว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าสูงขึ้นในทั้งสามวิธี และพิจารณาที่ขนาดตัวอย่าง 60, 100 และ 200 เมื่อร้อยละการสูญหายเพิ่มขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำลงในวิธีระยะห่างใกล้สุด ในขณะที่วิธีการถดถอยป๊วซงและวิธีถ่วงน้ำหนักเมื่อร้อยละการสูญหายเพิ่มขึ้นพบว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าสูงขึ้น ในกรณีวิธีการถดถอยป๊วซงจึงเป็นวิธีที่เหมาะสมที่สุดที่ในกรณีที่ขนาดตัวอย่าง 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่าง 20, 60 และ 100

พารามิเตอร์ชุดที่ 4 ในทุกขนาดตัวอย่าง พบว่าวิธีการถดถอยป๊วซงให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด รองลงมาคือวิธีถ่วงน้ำหนักและวิธีระยะห่างใกล้สุดตามลำดับ เมื่อพิจารณาในทุกวิธีพบว่าเมื่อร้อยละการสูญหายเพิ่มขึ้น ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจะมีค่าสูงขึ้นตามไปด้วย และเมื่อขนาดตัวอย่างตั้งแต่ 60 หน่วยขึ้นไป ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจะมีค่าลดลง ในกรณีนี้วิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุด

พารามิเตอร์ชุดที่ 5 วิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด รองลงมาคือ วิธีถ่วงน้ำหนักและวิธีการถดถอยป๊วซงตามลำดับ เมื่อร้อยละการสูญหายเพิ่มขึ้น วิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำลงไปด้วย พิจารณาที่ขนาดตัวอย่างเป็น 20 เมื่อค่าร้อยละการสูญหายเพิ่มขึ้นวิธีการถดถอยป๊วซงให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำลงและที่ขนาดตัวอย่างเป็น 60, 100 และ 200 ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยสูงขึ้น และเมื่อพิจารณาที่ขนาด

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตัวอย่างเป็น 20, 60 และ 100 เมื่อค่าร้อยละการสูญหายเพิ่มขึ้นวิธีถ่วงน้ำหนักจะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำลง แต่เมื่อขนาดตัวอย่างเป็น 200 วิธีนี้ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยขึ้นลงไม่สม่ำเสมอ ในกรณีนี้วิธีการถดถอยป๊วซงจึงเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 60, 100 และ 200 ส่วนวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุดเมื่อขนาดตัวอย่างเป็น 20

5.2 อภิปรายผลการศึกษา

จากผลการศึกษาการเปรียบเทียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายในการถดถอยป๊วซง พบว่าในชุดพารามิเตอร์ที่แตกต่างกันให้ผลที่แตกต่างกัน รวมทั้งเมื่อขนาดตัวอย่าง และร้อยละการสูญหายต่างกัน จะให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่างกันด้วย ในพารามิเตอร์ชุดที่ 1 2 และ 3 พบว่าวิธีระยะห่างใกล้สุดให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดแต่เนื่องจากเมื่อขนาดตัวอย่างที่มากขึ้นจะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ขึ้นลงไม่สม่ำเสมอ และเมื่อร้อยละการสูญหายเพิ่มมากขึ้นกลับให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำลง สอดคล้องกับ งานวิจัยของภาคย์ สิทธิผกาผล (2555) พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นและร้อยละของข้อมูลที่สูญหายเพิ่มขึ้น ค่าสัมประสิทธิ์การแปรผันมีแนวโน้มลดลง ดังนั้นจึงควรใช้ข้อมูลที่มีขนาดใหญ่และมีระดับของข้อมูลที่สูญหายอย่างเหมาะสมเพื่อช่วยลดความคลาดเคลื่อนกำลังสองเฉลี่ยและเพิ่มความแม่นยำในการประมาณ เพราะฉะนั้นวิธีระยะห่างใกล้สุดอาจจะไม่เหมาะสมในการนำมาให้ในกรณีนี้ควรหลีกเลี่ยงไปใช้วิธีการถดถอยป๊วซงหรือวิธีถ่วงน้ำหนัก ตามความเหมาะสมของขนาดตัวอย่างและร้อยละการสูญหาย

พารามิเตอร์ชุดที่ 4 ในทุกขนาดตัวอย่าง พบว่าวิธีการถดถอยป๊วซงให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด เพราะฉะนั้นวิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุด รองลงมาคือวิธีถ่วงน้ำหนักและวิธีระยะห่างใกล้สุดตามลำดับ เมื่อพิจารณาในทุกวิธีพบว่าเมื่อร้อยละการสูญหายเพิ่มขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจะมีค่าสูงขึ้นตามไปด้วย และเมื่อขนาดตัวอย่างตั้งแต่ 60 หน่วยขึ้นไป ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจะมีค่าลดลงสอดคล้องกับ Fatma El Zahra S. Salama, Ahmed M. Gad และ Amany M. Mohamed (2016) ได้ทำการศึกษาวิธีการวิเคราะห์เปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับข้อมูลจำนวนนับแบบระยะยาวที่มีข้อมูลสูญหาย ภายใต้การแจกแจงป๊วซง โดยศึกษาจำลองเพื่อเปรียบเทียบวิธีการต่างๆ ภายใต้สถานการณ์ที่แตกต่างกัน ในกลุ่ม MCAR จะเห็นได้ว่าวิธีการประมาณทั้งหมดสร้างการประมาณที่ค่อนข้างเป็นกลาง ความเอนเอียงสัมพัทธ์ต่ำกว่า 30 % สำหรับวิธีการทั้งหมด เมื่ออัตราการสูญหายและขนาดตัวอย่างที่เปลี่ยนแปลงไป ภายใต้ MCAR วิธี WGEE จะเอนเอียงน้อยกว่า และภายใต้ MAR วิธี WGEE นั้นดีกว่าในหลายกรณีแสดงว่าร้อยละของการสูญหายมีผลต่อวิธีการประมาณค่าสูญหาย แสดงให้เห็นว่าในวิธีการประมาณค่าและร้อยละการสูญหายต่างกัน ให้ผลที่แตกต่างกันด้วย

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

พารามิเตอร์ชุดที่ 5 ขนาดตัวอย่างเป็น 60, 100 และ 200 วิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด เมื่อร้อยละการสูญหายเพิ่มขึ้นวิธีระยะห่างใกล้สุดให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำลงไปด้วย และในวิธีการถดถอยป๊วซงและวิธีถ่วงน้ำหนักที่ขนาดตัวอย่างต่างๆ ทั้งสองวิธีจะให้ผลที่แตกต่างกันด้วยกันไป ซึ่งไม่สามารถบอกสรุปวิธีใดเป็นวิธีที่ดีที่สุดและพิจารณาที่ขนาดตัวอย่างต่างๆ ในทุกวิธีพบว่าเมื่อขนาดตัวอย่างสูงขึ้นกลับให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่สูงขึ้นตามไปด้วยสอดคล้องกับอุษณีย์ วงศ์อามาตย์(2555) ได้ทำการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงเส้น พบว่าวิธีการประมาณทุกวิธีสามารถประมาณได้ดีขึ้นเมื่อขนาดตัวอย่างมีขนาดใหญ่ขึ้น วิธีการประมาณทุกวิธีประมาณได้แย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย สัดส่วนของการสูญหาย และระดับของการสูญหายแบบอิกนอร์เรเบิล

5.3 ข้อเสนอแนะ

จากการศึกษาเปรียบเทียบวิธีการประมาณค่าสำหรับข้อมูลสูญหายภายใต้การแจกแจงป๊วซง ทั้ง 3 วิธี เมื่อค่าพารามิเตอร์ทุกตัวมีค่าเป็นบวก ในกรณีที่ขนาดตัวอย่างมีขนาดใหญ่วิธีการถดถอยป๊วซงเป็นวิธีที่เหมาะสมที่สุด ในกรณีตัวอย่างมีขนาดเล็กวิธีถ่วงน้ำหนักเป็นวิธีที่เหมาะสมที่สุด ในกรณีที่พารามิเตอร์ทุกตัวมีค่าเป็นลบ วิธีประมาณค่าสูญหายด้วยการถดถอยป๊วซงเป็นวิธีที่เหมะสมที่สุด และกรณีที่พารามิเตอร์มีค่าทั้งบวกและลบวิธีการถดถอยป๊วซงและวิธีการถ่วงน้ำหนักทั้งสองวิธีนี้ให้ผลที่ใกล้เคียงกัน การนำวิธีการประมาณค่าสูญหายในการค้นคว้าแบบอิสระนี้ไปใช้ควรเลือกวิธีการประมาณค่าสูญหายให้เหมาะสมตามที่ได้กล่าวข้างต้น

ส่วนวิธีระยะห่างใกล้สุดเมื่อร้อยละการสูญหายเพิ่มขึ้นพบว่าค่าคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าลดลง และเมื่อขนาดตัวอย่างเพิ่มขึ้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีแนวโน้มขึ้นลงไม่สม่ำเสมอ ฉะนั้นไม่แนะนำให้ใช้วิธีนี้ เนื่องจากวิธีระยะห่างใกล้สุดใช้หลักการของการถดถอยทั่วไป การนำมาใช้ในกรณีที่ตัวแปรตอบสนองเป็นจำนวนนับ ทำให้ค่าเกิดความเอนเอียงและให้ผลที่คลาดเคลื่อนได้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

เอกสารอ้างอิง

[1] ชูเกียรติ ผุดพรมราช. แนวคิดเชิงทฤษฎีการวิเคราะห์การถดถอยป๊วซงและการประยุกต์ [วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต]. เชียงใหม่: มหาวิทยาลัยเชียงใหม่; 2547.

[2] ทัตดา หิรัญพต. (2554). การเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยวิธีค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก. วิทยาศาสตรบัณฑิต. มหาวิทยาลัยบูรพา

[3] นรุตม์ บุตรพลอย. การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล. Natinal Conference on Information Technology 2553;28:25-29

[4] พรสิน สุภวาลย์. การวิเคราะห์การถดถอย. กรุงเทพฯ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพระนคร; 2556

[5] เพียงออ ยีสา. (2551). การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น. วิทยานิพนธ์มหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

[6] ภาคย์ สิทธิผกาผล. (2555). การประมาณค่าที่เก็บไม่ได้ในการสำรวจตัวอย่าง. วิทยานิพนธ์มหาบัณฑิต สถาบันบัณฑิตพัฒนาบริหารศาสตร์.

[7] อุษณีย์ วงศ์อามาตย์. (2555). การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์การถดถอยเชิงเส้นพหุ. วิทยานิพนธ์มหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

[8] Laaksonen, S. (2000) Regression-Based Nearest Neighbour Hot Decking. Computation Statistics; 15(1), 65-71

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

รายการสิ่งเผยแพร่

1) ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี “ข้อมูลสูญหายและแนวทางการจัดการ” Data Management & Biostatistics Journal Vol.4 No.3, Year 2009 Page No.52-61.

2) พัชรีพร สนรักษา, พัชรี วงษ์เกษม และ คณินทร์ ธีภาพโอฬาร “การประมาณและการเปรียบเทียบตัวประมาณค่าเฉลี่ยในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหาย” วารสารวิทยาศาสตร์ มศว. ปีที่ 28 ฉบับที่ 1, 2555

3) Fatma El Zahraa S. Salama, Ahmed M. Gad and Amany M. Mohamed, “Modeling Longitudinal Count Data with Missing Values: AComparative Study,” Applied Mathematical Science, Vol. 2, 8 September 2016

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ภาคผนวก ก

รูปแบบการจำลองสถานการณ์ การประมาณค่าสูญหายทั้งสามวิธี ได้แก่ วิธี NNI PRI และ WNPR ด้วยการจำลองจากโปรแกรม R version 3.3.1

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

การจำลองข้อมูลการการเขียนโค้ดในโปรแกรม R version 3.3.1

```
######################## Setting of Values ##############################

########## Regression Coefficient ############

b0 <- -1.5

b1 <- 0.2

b2 <- 0.5

b3 <- 0.8



######### Simple Size ########################

n<-20



####### Probaility of missing data ###########

prob_missing_data=0.15



####### Loop of simulation ###################

L=1000



########## Setting of MSE Variable ############

MSE_PRI<-rep(NA,L)

MSE_NNI<-rep(NA,L)

MSE_WNPR<-rep(NA,L)
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
#####################################Start of Simulation ##################

for (tt in 1:L){


######Generate random variable for Poisson Regression ############

x1 <- runif(n)

x2 <- runif(n)

x3 <- runif(n)

lam <- exp( b0 + b1 * x1 + b2 * x2 + b3 * x3 )



y.pois <- rpois(n,lam)

POIS<- data.frame( y.pois, x1, x2, x3 )



########## function of MAR missingness in y

generate.md <- function( data, pos = 1, Z = 2, pmis = prob_missing_data,

strength = c(prob_missing_data,1-prob_missing_data ) )

{

total <- round( pmis * nrow(data) )

sm <- which( data[,Z] < mean( data[,Z] ) )

 gr <- which( data[,Z] > mean( data[,Z] ) )
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
sel.sm <- sample( sm, round( strength[1] * total ) )

sel.gr <- sample( gr, round( strength[2] * total ) )

sel <- c( sel.sm, sel.gr )

data[sel,pos] <- NA

return(data)

}



########## MAR missingness to simulated data using generate.md function

MissingPOIS <- generate.md( POIS, pmis = prob_missing_data, strength = c( prob_missing_data,1-
prob_missing_data) )

new_MissingPOIS<- data.frame(MissingPOIS[is.na(MissingPOIS$y.pois),])

PRI=MissingPOIS

WNPR=MissingPOIS
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
##########Nearest-Neighbor Hot-Deck Imputation Method of estimating missing data #####

################### "eukl", Euclidean distance ######################

NNI=impute.NN_HD(DATA=MissingPOIS,distance="eukl",comp="mean",donor_limit=1,optimal
_donor="odd")

############End of Nearest-Neighbor Hot-Deck Imputation Method#########



####################### Poisson Regression Imputation ############################
```

```
A=glm( y.pois ~ x1 + x2 + x3, family = "poisson" ,data=MissingPOIS,na.action = "na.omit")


##### Selected Missing Data ######

new_PRI <- data.frame(PRI[is.na(PRI$y.pois),])

new_PRI

##### Predict Missing Data ######

predictlamda=exp(predict(A,new_PRI))


##### Random Missing Data function ######

mipois<-function(predictlamda){

data<-numeric()

                for(i in 1 : length(predictlamda))

                {

            a=predictlamda[i]

                        data[i]<- rpois(1, a)

                }

return(data)

}


##### Substance Missing Data  ######
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
BB=mipois(predictlamda)

PRI$y.pois[is.na(PRI$y.pois)]= BB

BB

##################End of Poisson Regression Imputation####################


########## Weighted Nearest Neighbor and Poisson Regression (WNPR) ##############


NNI1<- data.frame( MissingPOIS$y.pois,NNI)

NNI2<- data.frame(NNI1[is.na(NNI1$MissingPOIS.y.pois),])

NNI2$y.pois

CC<-rep(NA,length(new_MissingPOIS))

for(i in 1 : length(new_MissingPOIS))

{

        u<-runif(1)

        if (u<0.5) { CC[i]=BB[i] } else {CC[i]=NNI2$y.pois[i]}

}

WNPR$y.pois[is.na(WNPR$y.pois)]= CC


###########End of Weighted Nearest Neighbor and Poisson Regression (WNPR)########## #


#### Estimate Parameter for Poission regression with no estimating missing data ###########
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
respoisPRI<-glm( y.pois ~ x1 + x2 + x3, family = "poisson" ,data=PRI)

MSE_PRI[tt]<-var(respoisPRI$residuals)



respoisNNI<-glm( y.pois ~ x1 + x2 + x3, family = "poisson" ,data=NNI)

MSE_NNI[tt]<-var(respoisNNI$residuals)



respoisWNPR<-glm( y.pois ~ x1 + x2 + x3, family = "poisson" ,data=WNPR)

MSE_WNPR[tt]<-var(respoisWNPR$residuals)



}

########################  End of Simulation  ################################



############################## Report ###################################

MsePRI<-mean(MSE_PRI)

MseNNI<-mean(MSE_NNI)

MseWNPR<-mean(MSE_WNPR)

MsePRI

MseNNI

MseWNPR
```

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ประวัติผู้เขียน

ชื่อ-นามสกุล นางสาวศุภรัสมิ์ แย้มมีเชาว์

วัน เดือน ปี เกิด 7 พฤศจิกายน 2535

ประวัติการศึกษา ปีการศึกษา 2557 วิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์
มหาวิทยาลัยราชภัฏสวนสุนันทา

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved