การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง

โดย

นางสาวนลัทพร รูปหมอก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

สาขาวิชาสถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต

ภาควิชาสถิติ

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2562

การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง

โดย
นางสาวนลัทพร รูปหมอก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาสถิติ
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2562

A COMPARISON OF MISSING DATA IMPUTATION METHODS IN WITHIN-SUBJECT REPEATED MEASURE DESIGN

By

MISS Nalattaporn ROOPMOK

A Thesis Submitted in Partial Fulfillment of the Requirements

for Master of Science (APPLIED STATISTICS)

Department of STATISTICS

Graduate School, Silpakorn University

Academic Year 2019

หัวข้อ	การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง
โดย	นลัทพร รูปหมอก
สาขาวิชา	สถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. กมลชนก พานิชการ

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร.จุไรรัตน์ นันทานิช)

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ประหยัด แสงงาม)

.....อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร.กมลชนก พานิชการ)

.....ผู้ทรงคุณวุฒิภายนอก
(รองศาสตราจารย์ ดร.บุญอ้อม โฉมที)

60304201 : สถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
คำสำคัญ : ข้อมูลสูญหาย, แผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลอง, แทนที่ด้วยค่าเฉลี่ย, CopyMean, โครงข่ายประสาทเทียม

นางสาว นลัทพร รูปหมอก: การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง อาจารย์ที่ปรึกษาวิทยานิพนธ์ : ผู้ช่วยศาสตราจารย์ ดร. กมลชนก พานิชการ

แผนแบบการทดลองแบบวัดซ้ำมีลักษณะการเก็บข้อมูลจากหน่วยตัวอย่างเดียวกันแต่ต่างกันที่ช่วงเวลาหรือเงื่อนไขอื่น ซึ่งนิยมใช้ในงานวิจัยทางด้านการแพทย์หรือสาธารณสุข งานวิจัยนี้เสนอการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองเมื่อสุ่มค่าข้อมูลสูญหายอย่างสุ่มสมบูรณ์ โดยประยุกต์จากวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean Trajectory วิธี CopyMean LOCF และวิธีโครงข่ายประสาทเทียม โดยใช้เกณฑ์ในการประเมินด้วยค่า MAD RMSD และค่า Bias ซึ่งทำการทดลองทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลองโดยในชุดข้อมูลจำลองกำหนดให้ในแต่ละตัวแปรมีค่าเฉลี่ยและค่าความแปรปรวนเท่ากัน ผลการวิจัยพบว่าในกรณีส่วนใหญ่วิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายในข้อมูลจริงและข้อมูลจำลองในกรณีไม่มีสหสัมพันธ์และสหสัมพันธ์น้อย (0 0.3 และ 0.5) ส่วนในข้อมูลจำลองกรณีที่สหสัมพันธ์ค่อนข้างมาก (0.7 และ 0.9) วิธี CopyMean Trajectory เป็นวิธีการที่ดีที่สุดในกรณีส่วนใหญ่

60304201 : Major (APPLIED STATISTICS)
Keyword : missing values, Within-subject repeated measure design, Mean Substitution, CopyMean, Artificial Neural Network

MISS NALATTAPORN ROOPMOK : A COMPARISON OF MISSING DATA IMPUTATION METHODS IN WITHIN- SUBJECT REPEATED MEASURE DESIGN THESIS ADVISOR : ASSISTANT PROFESSOR KAMOLCHANOK PANISHKAN, Ph.D.

Within-subject repeated measure design is an experimental design which has the characteristic of collecting data from the same sample unit at different times or other conditions. It is popular in medical or public health research. This research presents a comparison of missing data imputation methods in within-subject repeated measure design when missing values are Missing Completely at Random. The imputation methods were applied by the Mean Substitution method, CopyMean Trajectory method, CopyMean LOCF method and Artificial Neural Network method by using 3 assessment criteria such as MAD RMSD and Bias. All these methods were tested on both real dataset and artificial datasets when defined mean and variance are equal in each variable. The results showed, in the most cases, the artificial neural network method performs the best in real dataset and in artificial datasets with no correlation or low correlation (0, 0. 3 and 0. 5) . However, in artificial datasets with high correlation (0.7 and 0.9), the CopyMean Trajectory method is the best method in the most cases.

กิตติกรรมประกาศ

งายวิจัยนี้สำเร็จลุล่วงได้ด้วยความกรุณาช่วยเหลือ แนะนำ ให้คำปรึกษา ตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่อย่างดียิ่งจาก ผศ. ดร. กมลชนก พานิชการ ที่ปรึกษาหลักของงานวิจัยนี้ ขอขอบพระคุณ ผศ. ดร. ประหยัด แสงงาม และ รศ. ดร. บุญอ้อม โฉมที ที่กรุณาช่วยตรวจสอบแก้ไขและให้คำปรึกษาเกี่ยวกับงานวิจัย ผู้เขียนกราบขอบพระคุณเป็นอย่างสูง รวมไปถึงทุนเรียนดีวิทยาศาสตร์แห่งประเทศไทยที่ได้สนับสนุนทุนทรัพย์ในการเรียนของข้าพเจ้าจนสำเร็จการศึกษา และขอขอบคุณญาติพี่น้องทุกคนที่ช่วยเหลือสนับสนุนทั้งด้านกำลังใจและกำลังทรัพย์ด้วยดีตลอดมา นอกจากนี้ยังมีผู้ที่ให้ความร่วมมือช่วยเหลืออีกหลายท่าน ซึ่งผู้วิจัยไม่สามารถกล่าวนามในที่นี้ได้หมด จึงขอขอบคุณทุกท่านเหล่านั้นไว้ ณ โอกาสนี้ด้วย

นลัทพร รูปหมอก

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ญ

สารบัญรูปภาพ ..... ฑ

บทที่ 1 บทนำ 1

1. ความเป็นมาและความสำคัญของปัญหา 1
2. วัตถุประสงค์ของการวิจัย 5
3. สมมติฐานของการวิจัย 5
4. ขอบเขตของการวิจัย 5
5. นิยามศัพท์เฉพาะ 7
6. ประโยชน์ที่คาดว่าจะได้รับ 7

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 8

1. ทฤษฎีที่เกี่ยวข้อง 8
    - 1.1 ข้อมูลตามคาบเวลา (Longitudinal data) 8
    - 1.2 ข้อมูลแบบวัดซ้ำ (Repeated Measure data) 9
    - 1.3 ชุดข้อมูลจริงที่ใช้ในงานวิจัย 15
    - 1.4 โครงข่ายประสาทเทียม (Artificial Neural Network) 18
    - 1.5 การประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS) 26

1.6 การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean 29
1.7 การประมาณค่าข้อมูลสูญหายด้วยวิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN) 34
1.8 เกณฑ์ที่ใช้ในการประเมิน 40
1.8.1 ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute deviation : MAD) 40
1.8.2 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square deviation : RMSD) 42
1.8.3 ค่าความเอนเอียง (Bias) 43
2. งานวิจัยที่เกี่ยวข้อง 45
บทที่ 3 วิธีดำเนินงานวิจัย 49
1. ข้อมูลที่ใช้ในการวิจัย 49
2. เครื่องมือที่ใช้ในการวิจัย 49
3. สถิติที่ใช้ในการวิจัย 50
4. ขั้นตอนการจำลองข้อมูล 50
5 วิธีการวิเคราะห์ข้อมูล 52
บทที่ 4 ผลการวิจัย 55
ส่วนที่ 1 ผลการวิจัยโดยใช้ชุดข้อมูลจริง 55
1.1 ข้อมูลชุด Drug Effect 55
1.2 ข้อมูลชุด Skydrive 58
1.3 ข้อมูลชุด Fecal Fat 61
ส่วนที่ 2 ผลการวิจัยโดยใช้ชุดข้อมูลจำลอง 64
2.1 ชุดข้อมูลจำลองที่ 1 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0) 65
2.2 ชุดข้อมูลจำลองที่ 2 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.3) 69
2.3 ชุดข้อมูลจำลองที่ 3 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.5) 73

2.4 ชุดข้อมูลจำลองที่ 4 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.7)..... 77

2.5 ชุดข้อมูลจำลองที่ 5 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9)..... 81

บทที่ 5 สรุป อภิปรายผล และข้อเสนอแนะ 86

1. สรุปผลการวิจัย..... 86

2. อภิปรายผลการวิจัย..... 90

3. ข้อเสนอแนะ..... 91

รายการอ้างอิง..... 92

ภาคผนวก..... 95

โปรแกรมคอมพิวเตอร์ที่ใช้ในงานวิจัย..... 96

ประวัติผู้เขียน..... 106

สารบัญตาราง

หน้า

ตารางที่ 1 รูปแบบของข้อมูลแบบวัดซ้ำ 10

ตารางที่ 2 ตารางวิเคราะห์ความแปรปรวน 12

ตารางที่ 3 ข้อมูลของตัวอย่างที่ 1 ข้อมูลแบบวัดซ้ำ 13

ตารางที่ 4 ตารางวิเคราะห์ความแปรปรวนสำหรับตัวอย่างที่ 1 14

ตารางที่ 5 ข้อมูลชุด Drug Effect 15

ตารางที่ 6 ข้อมูลชุด Skydive 17

ตารางที่ 7 ข้อมูลชุด Fecal Fat 18

ตารางที่ 8 ตัวอย่างที่ 2 ข้อมูลน้ำหนักของผมหลังใช้ยา Minoxidil 27

ตารางที่ 9 ข้อมุลสูญหายจากตัวอย่างที่ 2 27

ตารางที่ 10 ค่าเฉลี่ยตามขวางของตัวอย่างที่ 2 ในตารางที่ 6 28

ตารางที่ 11 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย 29

ตารางที่ 12 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ Trajectory Mean..... 31

ตารางที่ 13 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory 32

ตารางที่ 14 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ LOCF 33

ตารางที่ 15 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF..... 34

ตารางที่ 16 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียม 39

ตารางที่ 17 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีCopyMean และวิธีโครงข่ายประสาทเทียม 40

ตารางที่ 18 ค่า MAD สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม 41

ตารางที่ 19 ค่า RMSD สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม 43

ตารางที่ 20 ค่า Bias สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม 44

ตารางที่ 21 ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายสำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธี วิธีการแทนที่ด้วยค่าเฉลี่ย วิธีCopyMean และวิธีโครงข่ายประสาทเทียม 45

ตารางที่ 22 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Drug Effect 56

ตารางที่ 23 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 1 ค่า 56

ตารางที่ 24 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 2 ค่า 57

ตารางที่ 25 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 3 ค่า 57

ตารางที่ 26 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Skydrive 59

ตารางที่ 27 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 1 ค่า 59

ตารางที่ 28 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 2 ค่า 60

ตารางที่ 29 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 3 ค่า 60

ตารางที่ 30 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Fecal Fat 61

ตารางที่ 31 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 1 ค่า 62

ตารางที่ 32 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 2 ค่า 62

ตารางที่ 33 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 3 ค่า 63

ตารางที่ 34 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 1 ค่า 65

ตารางที่ 35 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 2 ค่า 66

ตารางที่ 36 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 3 ค่า 68

ตารางที่ 37 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า 69

ตารางที่ 38 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 2 ค่า 70

ตารางที่ 39 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 3 ค่า 72

ตารางที่ 40 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 1 ค่า 73

ตารางที่ 41 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 2 ค่า 74

ตารางที่ 42 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 3 ค่า 76

ตารางที่ 43 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 1 ค่า 77

ตารางที่ 44 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 2 ค่า 78

ตารางที่ 45 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 3 ค่า 80

ตารางที่ 46 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 1 ค่า 81

ตารางที่ 47 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 2 ค่า 82

ตารางที่ 48 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 3 ค่า 84

ตารางที่ 49 ผลการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจริง 87

ตารางที่ 50 ผลการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลอง..... 89

สารบัญรูปภาพ

หน้า

ภาพที่ 1 แผนภาพกระบวนการทำงานของโครงข่ายประสาทเทียม 19

ภาพที่ 2 กราฟของฟังก์ชันขั้นบันได 22

ภาพที่ 3 กราฟของฟังก์ชันเชิงเส้น 22

ภาพที่ 4กราฟของฟังก์ชันเส้นโค้งซิกมอยด์ 23

ภาพที่ 5 กราฟของฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ 24

ภาพที่ 6 กราฟของ ReLU Function 25

ภาพที่ 7 กราฟของฟังก์ชัน Swish 25

ภาพที่ 8 กระบวนการ backpropagation 35

ภาพที่ 9 แผนภาพโครงข่ายประสาทเทียมสำหรับการประมาณค่าข้อมูลสูญหายของ y_{22} 38

ภาพที่ 10แผนภาพโครงข่ายประสาทเทียมสำหรับการประมาณค่าข้อมูลสูญหายของ y_{33} 39

ภาพที่ 11 แผนภาพแสดงลำดับขั้นตอนการทำงานบนชุดข้อมูลจริง 53

ภาพที่ 12 แผนภาพแสดงลำดับขั้นตอนการทำงานบนชุดข้อมูลจำลอง 54

ภาพที่ 13 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 1 ค่า 66

ภาพที่ 14 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 2 ค่า 67

ภาพที่ 15 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 3 ค่า 68

ภาพที่ 16 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า 70

ภาพที่ 17 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 2 ค่า 71

ภาพที่ 18 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 3 ค่า 72

ภาพที่ 19 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 1 ค่า 74

ภาพที่ 20 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาp ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 2 ค่า 75

ภาพที่ 21 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 3 ค่า 76

ภาพที่ 22 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 1 ค่า 78

ภาพที่ 23 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 2 ค่า 79

ภาพที่ 24 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 3 ค่า 80

ภาพที่ 25 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 1 ค่า 82

ภาพที่ 26 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 2 ค่า 83

ภาพที่ 27 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 3 ค่า 84

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ค่าข้อมูลสูญหายคือค่าในชุดข้อมูลที่ค่าสังเกตที่สนใจหายไปในแต่ละตัว โดยข้อมูลสูญหายนั้นนับเป็นปัญหาสำคัญที่อาจส่งผลให้เกิดปัญหาหลายประการเช่น ข้อมูลสูญหายทำให้กำลังการทดสอบลดลงกล่าวคือจะลดความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่างเมื่อสมมติฐานว่างไม่เป็นจริง ทำให้การประมาณค่าพารามิเตอร์เกิดความเอนเอียง ทำให้เห็นสารสนเทศจากตัวอย่างลดลง และทำให้การวิเคราะห์ข้อมูลซับซ้อนขึ้นเป็นต้น(Kang, 2013) ในโปรแกรมสำเร็จรูปทางสถิติส่วนใหญ่จะใช้วิธีการตัดค่าสังเกตที่มีค่าสูญหายตั้งแต่ 1 ค่าขึ้นไปออกจากการวิเคราะห์ซึ่งเป็นผลให้ขนาดตัวอย่างลดลงตามไปด้วย (Vermeulen et al., 2005) ข้อมูลสูญหายจะมีผลกระทบอย่างมากในแผนแบบการทดลอง เพราะข้อมูลสูญหายในแผนแบบการทดลองนั้นจะทำให้การวิเคราะห์ผลในบางแผนแบบทำได้ยากหรืออาจทำไม่ได้เลย ลักษณะของข้อมูลสูญหายแบ่งออกได้เป็น 3 ประเภท ประเภทของการสูญหายแบบแรกคือการสูญหายอย่างสุ่มสมบูรณ์ (Missing Completely at Random : MCAR) มีลักษณะสำคัญคือความน่าจะเป็นของการเกิดค่าข้อมูลสูญหายเป็นอิสระกับทั้งค่าของข้อมูลที่ถูกเก็บมาแล้วและค่าของข้อมูลที่ยังไม่ได้เก็บมา ประเภทของการสูญหายแบบที่สองคือการสูญหายอย่างสุ่ม (Missing at Random : MAR) มีลักษณะคือความน่าจะเป็นของการเกิดค่าข้อมูลสูญหายมีความเกี่ยวพันกับค่าบางค่าของข้อมูลที่ถูกเก็บมาแล้วแต่เป็นอิสระกับค่าของข้อมูลที่ยังไม่ได้เก็บมา และประเภทของการสูญหายแบบสุดท้ายคือ การสูญหายแบบไม่สุ่ม (Missing Not at Random : MNAR) มีลักษณะคือความน่าจะเป็นของการเกิดค่าข้อมูลสูญหายมีความเกี่ยวพันกับทั้งค่าของข้อมูลที่ถูกเก็บมาแล้วและค่าของข้อมูลที่ยังไม่ได้เก็บมา (Little & Rubin, 1987) การแก้ปัญหาเกี่ยวกับข้อมูลสูญหายโดยทั่วไปอาจทำได้โดยการตัดค่าสังเกตที่มีข้อมูลสูญหายออกเช่นวิธีการที่เรียกว่า Listwise Deletion หรือ Case Deletion เป็นการตัดค่าสังเกตที่มีค่าข้อมูลสูญหายออกทั้งค่าสังเกต และ Pairwise Deletion ที่เป็นการตัดเฉพาะส่วนที่เป็นค่าข้อมูลสูญหายออกแล้ววิเคราะห์เฉพาะข้อมูลที่เหลืออยู่(Lani 2019) หรืออาจใช้วิธีประมาณค่าข้อมูลสูญหายก็ได้ การตัดค่าสังเกตที่มีข้อมูลสูญหายออกซึ่งโดยทั่วไปแล้วการใช้โปรแกรมสำเร็จรูปทางสถิติจะใช้วิธีการนี้ในการวิเคราะห์ผล ซึ่งวิธีการนี้จะไม่เกิดปัญหาหากค่าสูญหายเป็นค่าที่ไม่มีความแตกต่างจากค่าสังเกตอื่นในชุดข้อมูลนั้นมากนักหรือค่าสูญหายเป็นค่าผิดปรกติ วิธีการตัดข้อมูลสูญหายออกนั้นก็อาจก่อให้เกิดปัญหาตามที่กล่าวไปข้างต้น อย่างไรก็ตามการตัดค่าข้อมูลสูญหายออกเป็นการทำให้ชุดข้อมูลไม่สมบูรณ์ดังนั้นการประมาณค่าข้อมูลสูญหายจึงเป็นสิ่งจำเป็นเพื่อทำให้การวิเคราะห์ข้อมูลเกิดความถูกต้องมากขึ้น

วิธีการที่จะประมาณค่าข้อมูลสูญหายที่นิยมใช้ทั่วไปมีหลายวิธีเช่น วิธีแทนที่ด้วยค่าเฉลี่ย (Mean Substitution) คือเป็นการแทนที่ค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณด้วยการวิเคราะห์การถดถอย (Regression Imputation) คือ วิธีการที่กำหนดให้ตัวแปรที่มีค่าข้อมูลสูญหายเป็นตัวแปรตามและใช้ตัวแปรอื่นที่เหลือเป็นตัวแปรต้นเพื่อพยากรณ์ค่าข้อมูลสูญหาย วิธีการแทนที่ด้วยการวัดความรู้ครั้งล่าสุด (Last Observation Carried Forward : LOCF) คือวิธีการแทนที่ค่าข้อมูลสูญหายด้วยข้อมูล ก่อนหน้าที่ไม่เป็นค่าข้อมูลสูญหาย (Genolini และคณะ, 2016) วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood : ML) คือวิธีการแทนที่ค่าข้อมูลสูญหายด้วยการคำนวณจากวิธีภาวะน่าจะเป็นสูงสุด(Kang, 2013) หรือวิธีค่าคาดหมายสูงสุด (Expectation Maximization : EM) คือวิธีการประมาณค่าข้อมูลสูญหายด้วยหลักการของวิธีภาวะน่าจะเป็นสูงสุดแต่ใช้วิธีประมาณค่าพารามิเตอร์ 2 ขั้นตอน ได้แก่ ขั้นตอนประมาณค่าคาดหวัง (Expectation) เรียกว่า E step เป็นการประมาณลอการิทึมของฟังก์ชันภาวะน่าจะเป็นของฟังก์ชันพารามิเตอร์ และขั้นตอนการหาค่าสูงสุด (Maximization) เรียกว่า M step เป็นขั้นตอนการแทนที่ค่าข้อมูลสูญหายด้วยค่าคาดหวังแล้วประมาณค่าคาดหวังตัวใหม่ที่ได้ค่าไม่เปลี่ยนแปลงไปจากเดิมหรือเปลี่ยนแปลงในขนาดที่ยอมรับได้เป็นต้น(Rubin, Witkiewitz, Andre, & Reilly, 2007)

ข้อมูลแต่ละชนิดเหมาะกับวิธีการประมาณค่าข้อมูลสูญหายที่แตกต่างกัน Bingham, Stemmler, Peterson และ Graber (1998) ได้ทำงานวิจัยเพื่อศึกษาการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำ โดย Bingham และคณะได้ใช้วิธีการสุ่มค่าข้อมูลสูญหาย ออกเป็นสองส่วน ส่วนแรกคือการจำลองแบบครั้งเดียว คือจะสุ่มข้อมูลโดยใช้ค่าพารามิเตอร์จากชุดข้อมูลจริงจำนวน 183 ชุดข้อมูล แต่ละชุดข้อมูลสุ่มตัวอย่างขนาด 5 หน่วยตัวอย่างและวัดซ้ำ 4 ครั้ง โดยมีลักษณะของข้อมูล คือเป็นกราฟเส้นตรง กราฟกำลัง พาราโบลิก กราฟรูปตัวเอส และกราฟไซน์ตามลำดับ ส่วนที่สองจะใช้วิธีการจำลองข้อมูลแบบมอนติคาร์โลโดยกำหนดจำนวนการวนซ้ำ 1000 รอบ โดยกำหนดรูปร่างของข้อมูลคือกราฟเส้นตรง กราฟกำลังตามแนวตั้ง กราฟกำลังตามแนวนอน พาราโบลิก กราฟรูปตัวเอส และกราฟไซน์ แต่ละฟังก์ชั่นจะสุ่มขนาดตัวอย่างจำนวน 120, 240 และ 480 ค่าโดยแต่ละชุดข้อมูลสุ่มตัวอย่างขนาด 5 หน่วยตัวอย่างและวัดซ้ำ 4 ครั้งในแต่ละหน่วยตัวอย่างและสุ่มค่าข้อมูลสูญหาย จากนั้นประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยและเปรียบเทียบกับข้อมูลที่ไม่มีค่าข้อมูลสูญหายด้วยกราฟ Q-Q plot และทดสอบความแตกต่างของค่าโมเมนต์ด้วยสถิติทดสอบไคสแควร์และเปรียบเทียบค่าเฉลี่ยและค่าความคลาดเคลื่อนของข้อมูลแต่ละชุด ปรากฏว่าค่าประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าโมเมนต์ใกล้เคียงกับค่าจริง แต่เมื่อสัดส่วนของข้อมูลสูญหายเพิ่มความคลาดเคลื่อนในการประมาณค่าก็จะเพิ่มไปด้วยและการประมาณค่าข้อมูลสูญหายไม่ได้มีผลแตกต่างกันตามรูปร่างของข้อมูล

นอกจากนี้ Genolini, Lacombe, cochard, และ Subtil (2016) ได้ศึกษาวิธีการประมาณค่าข้อมูลสูญหายที่เรียกว่าวิธี CopyMean ซึ่งเป็นวิธีการใหม่ที่ใช้ในการพยากรณ์ค่าข้อมูลสูญหายแบบทางเดียวในการศึกษาตามคาบเวลา (Longitudinal Study) โดย Genolini, Lacombe, cochard, และ Subtil สนใจการประมาณค่าข้อมูลสูญหายด้วยกระบวนการประมาณค่าครั้งเดียว (Single Imputation) โดยไม่พิจารณาตัวแบบหรือวิธีการที่ใช้พื้นฐานของภาวะน่าจะเป็นเนื่องจากเขาได้พิจารณาว่าข้อมูลตามธรรมชาติส่วนใหญ่แล้วจะเป็นข้อมูลแบบไม่ใช้พารามิเตอร์แต่อาจมีลักษณะข้อมูลที่เปลี่ยนแปลงไปตามช่วงเวลา ซึ่งต้องการใช้สารสนเทศของข้อมูลก่อนหน้ามากกว่าลักษณะการแจกแจงของข้อมูลในการประมาณค่าข้อมูลสูญหาย หรือในบางกรณีวิธีการทางสถิติจะเหมาะสมเฉพาะในข้อมูลที่ไม่มีค่าสูญหาย โดยในกรณีนี้การประมาณค่าข้อมูลสูญหายจะไม่เหมาะสม ซึ่งวิธีการที่เขาสนใจที่จะใช้ในการประมาณค่าข้อมูลสูญหายในงานวิจัยนี้แบ่งออกเป็น 3 วิธีการได้แก่ วิธีการที่ 1 เป็นการประมาณค่าตามขวาง (Cross-sectional Imputation) คือวิธีการประมาณค่าข้อมูลสูญหายโดยใช้ข้อมูลอันเนื่องมาจากเวลามาใช้ในการประมาณค่าข้อมูลสูญหายโดยใช้การประมาณค่าด้วยวิธี ค่าเฉลี่ยตามขวาง (Cross Mean) ค่ามัธยฐานตามขวาง (Cross Median) ค่าสุ่มตามขวาง (Cross Hot Deck) วิธีการที่ 2 เป็นการประมาณค่าตามคาบเวลา(Longitudinal Imputation) คือวิธีการประมาณค่าข้อมูลสูญหายโดยใช้ข้อมูลที่ไม่เป็นค่าสูญหายจากตัวอย่างเดียวกันมาใช้ในการประมาณค่าข้อมูลสูญหายโดยใช้การประมาณค่าด้วยวิธี ค่าเฉลี่ยตามคาบเวลา (Traj Mean) ค่ามัธยฐานตามคาบเวลา (Traj Median) ค่าสุ่มตามคาบเวลา (Traj Hot deck) LOCF การประมาณค่าในช่วงเชิงเส้นสัมบูรณ์ (Interpolation Global) การประมาณค่าในช่วงเชิงเส้นสัมพัทธ์ (Interpolation local) การประมาณค่าในช่วงเชิงเส้นแบ่งครึ่ง (Interpolation Bisector) การประมาณค่าในช่วงเชิงเส้นโค้ง (Spline Interpolation) และ และวิธีการสุดท้ายเป็นการผสมระหว่างกระบวนการประมาณค่าตามขวาง และกระบวนการประมาณค่าตามคาบเวลาคือกระบวนการประมาณค่าข้อมูลสูญหายโดยใช้กระบวนการของสองกระบวนการมารวมกันเป็นการใช้ข้อมูลทั้งจากตัวอย่างและเวลาในการประมาณค่าข้อมูลสูญหาย โดยใช้การประมาณค่าด้วยวิธีการวิเคราะห์การถดถอย และ CopyMean ผู้วิจัยใช้วิธีการเหล่านี้ในการประมาณค่าข้อมูลที่มีลักษณะการสูญหาย 3 ลักษณะคือ MCAR MAR และ MNAR จากทั้งข้อมูลจริงและการจำลองข้อมูล โดยใช้เกณฑ์การประเมินประสิทธิภาพของการประมาณค่าข้อมูลสูญหายด้วยวิธีการค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย(Mean Absolute Deviation : MAD) รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย(Root Mean Square Deviation : RMSD) ค่าความเอนเอียง (bias) และค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) เพื่อใช้ในการประเมินการประมาณค่าข้อมูลสูญหายจากการจำลองข้อมูล ซึ่งผลลัพธ์ปรากฏว่าวิธี CopyMean LOCF เป็นวิธีการที่เหมาะสมในเกือบทุกสถานการณ์

ข้อมูลตามคาบเวลาเป็นการเก็บข้อมูลจากตัวแปรเดิมซ้ำกันเป็นระยะเวลาที่ยาวนานโดยจะเป็นส่วนขยายมาจากข้อมูลแบบวัดซ้ำ (Repeated Data) ที่ใช้สำหรับการวิเคราะห์ข้อมูลด้วยการวิเคราะห์ความแปรปรวนแบบวัดซ้ำแต่ต่างกันที่ข้อมูลแบบวัดซ้ำจะเป็นการวัดซ้ำในช่วงเวลาที่สั้นกว่าเมื่อทำการวัดซ้ำ ดังนั้นหน่วยตัวอย่างจะไม่เป็นอิสระกันด้วยซึ่งข้อมูลสูญหายในข้อมูลลักษณะนี้ก็สามารถทำให้การวิเคราะห์ข้อมูลเกิดความผิดพลาดได้เช่นกันและอาจทำให้เห็นแนวโน้มของข้อมูลอย่างไม่ต่อเนื่องอีกด้วย ฉะนั้นจึงควรแก้ปัญหาข้อมูลสูญหาย และเนื่องจากข้อมูลแบบวัดซ้ำมีลักษณะคล้ายกับข้อมูลตามคาบเวลาดังนั้นผู้วิจัยจึงสนใจที่จะนำวิธีการประมาณค่าข้อมูลสูญหายที่ใช้ได้ดีในข้อมูลตามคาบเวลา คือวิธีการ CopyMean มาทดลองประมาณค่าข้อมูลสูญหายในข้อมูลแบบวัดซ้ำ และนอกจากวิธีการนี้คาดว่าวิธีการแทนที่ด้วยค่าเฉลี่ยที่เป็นวิธีการทั่วไปในการประมาณค่าข้อมูลสูญหายก็น่าจะนำมาใช้ในการประมาณค่าข้อมูลสูญหายในข้อมูลแบบวัดซ้ำได้เช่นกัน

ดังนั้นในงานวิจัยนี้จะพิจารณาวิธีการประมาณค่าข้อมูลสูญหาย 3 วิธีคือ วิธี CopyMean ซึ่งเป็นวิธีการที่ดีที่สุดจากงานวิจัยของ Genolini, Lacombe, cochard และ Subtil (2016) สำหรับข้อมูลตามคาบเวลา วิธีการแทนที่ด้วยค่าเฉลี่ยเป็นวิธีการที่ดีที่สุดจากงานวิจัยของ Bingham, Stemmler, Peterson และ Graber (1998) ซึ่งเป็นวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำ และวิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN) ซึ่งเป็นวิธีการที่นิยมใช้ในทางคอมพิวเตอร์ มาประมาณค่าข้อมูลสูญหายทั้งจากข้อมูลจริงและจากการจำลองข้อมูลและใช้เกณฑ์การประเมินประสิทธิภาพของวิธีการด้วยค่า MAD ค่า RMSD และค่าความเอนเอียงเพื่อวัดความสามารถในการพยากรณ์ข้อมูลในแต่ละสถานการณ์

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อศึกษาวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำได้แก่

2.1.1 วิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS)

2.1.2 วิธี CopyMean ประกอบด้วย 2 วิธีย่อยได้แก่

2.1.2.1 วิธี CopyMean Trajectory (CopyMean Trajectory : CT)

2.1.2.2 วิธี CopyMean LOCF (CopyMean LOCF :CL)

2.1.3 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN)

2.2 เพื่อเปรียบเทียบวิธีการในการประมาณค่าข้อมูลสูญหายโดยใช้เกณฑ์ประเมินได้แก่

2.2.1 ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD)

2.2.2 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Deviation : RMSD)

2.2.3 ค่าความเอนเอียง (Bias)

2.3 เพื่อหาวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพในแต่ละสถานการณ์

3. สมมติฐานของการวิจัย

วิธีการประมาณค่าข้อมูลสูญหายมีประสิทธิภาพแตกต่างกันในแต่ละสถานการณ์

4. ขอบเขตของการวิจัย

ในงานวิจัยนี้จะเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลอง โดยเปรียบเทียบวิธีการประมาณค่าคือ วิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS) วิธี CopyMean และ วิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN) โดยกำหนดขอบเขตของการศึกษาดังนี้

4.1 ศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS) วิธี CopyMean และ วิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN)

4.2 ใช้เกณฑ์ในการประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย(MAD) รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSD) และค่าความเอนเอียง(Bias)

4.3 ข้อมูลจริงที่ใช้ในการทดลองประมาณค่าข้อมูลสูญหายมี 3 ชุดได้แก่

4.3.1 ข้อมูลชุด Drug Effect(Winer, 1962)

4.3.2 ข้อมูลชุด Skydive (Singley, Hale, & Russell, 2012)

4.3.3 ข้อมูลชุด The Fecal Fat(Vittinghoff, Glidden, Shiboski, & McCulloch, 2012)

4.4 ข้อมูลจำลองในแต่ละสถานการณ์กำหนดเงื่อนไขดังนี้

4.4.1 กำหนดลักษณะของข้อมูลสูญหายเป็นแบบการสูญหายอย่างสุ่มสมบูรณ์ (Missing Completely at Random : MCAR)

4.4.2 จำนวนค่าข้อมูลสูญหายเท่ากับ 1, 2 และ 3 ค่าตามลำดับ

4.4.3 การแจกแจงของข้อมูลที่ศึกษามีการแจกแจงปรกติพหุ 4 ตัวแปร (k=4) ที่ขนาดตัวอย่าง 5 (n=5) ภายใต้พารามิเตอร์ที่กำหนดคือ

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} \text{ และ } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

เมื่อ $\sigma_{ij} = \rho_i \sigma_i \sigma_j ; i \neq j; i, j = 1,2,3,4$

4.4.3.1 ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์คือ

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20$$
$$\sigma_i^2 = 25; \; i = 1,2,3,4$$
$$\rho_{ij} = 0 \;\; ; i \neq j \;, i, j = 1,2,3,4$$

4.4.3.2 ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์คือ

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20$$
$$\sigma_i^2 = 25; \; i = 1,2,3,4$$
$$\rho_{ij} = 0.3 \;\; ; i \neq j \;, i, j = 1,2,3,4$$

4.4.3.3 ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์คือ

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.5\ ;i \neq j\ ,i,\ j = 1,2,3,4$$

4.4.3.4 ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์คือ

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.7\ ;i \neq j\ ,i,\ j = 1,2,3,4$$

4.4.3.5 ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์คือ

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.9\ ;i \neq j\ ,i,\ j = 1,2,3,4$$

5. นิยามศัพท์เฉพาะ

5.1 การศึกษาตามขวาง หมายถึงการศึกษาที่มีจุดประสงค์เพื่อวิเคราะห์ข้อมูลตามช่วงเวลาที่เก็บมาและเป็นการศึกษาที่จะเก็บข้อมูลหลาย ๆ ตัวแปรเพื่อที่จะได้สารสนเทศของในแต่ละช่วงเวลาอย่างครบถ้วน

5.2 การศึกษาตามคาบเวลา หมายถึงการศึกษาที่มีจุดประสงค์เพื่อวิเคราะห์ข้อมูลในแต่ละตัวแปร และเป็นการศึกษาที่จะเก็บข้อมูลในช่วงระยะเวลาที่ยาวนานเพื่อที่จะได้สารสนเทศของตัวแปรอย่างครบถ้วน

5.3 ข้อมูลแบบวัดซ้ำ หมายถึงข้อมูลที่วัดค่าจากตัวแปรเดิมซ้ำ ๆ กันในต่างช่วงเวลา

5.4 ข้อมูลตามคาบเวลา หมายถึงข้อมูลที่เก็บจากการศึกษาตามคาบเวลา

5.5 ค่าข้อมูลสูญหาย หมายถึงค่าในชุดข้อมูลที่ไม่สามารถเก็บได้ หรือไม่ปรากฏในชุดข้อมูล

6. ประโยชน์ที่คาดว่าจะได้รับ

6.1 ทราบวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพในแต่ละสถานการณ์

6.2 ทำให้การวิเคราะห์ข้อมูลที่มีค่าสูญหายถูกต้องมากขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในส่วนของทฤษฎีและงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้จะแบ่งออกเป็นสองส่วนดังนี้

1. ทฤษฎีที่เกี่ยวข้อง
   1.1 ข้อมูลตามคาบเวลา
   1.2 ข้อมูลแบบวัดซ้ำ
   1.3 ชุดข้อมูลจริงที่ใช้ในงานวิจัย
   1.4 โครงข่ายประสาทเทียม
   1.5 การประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย
   1.6 การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean
   1.7 การประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียม
   1.8 เกณฑ์ที่ใช้ในการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
2. งานวิจัยที่เกี่ยวข้อง

1. ทฤษฎีที่เกี่ยวข้อง

1.1 ข้อมูลตามคาบเวลา (Longitudinal data)

ข้อมูลตามคาบเวลา (Longitudinal data) คือข้อมูลที่สังเกตค่าของตัวแปรในช่วงระยะเวลาที่ยาวนานมักจะเก็บข้อมูลเป็นระยะเวลาเป็นปีหรือเป็นทศวรรษซึ่งการศึกษาตามยาวนั้นจะทำให้เห็นสภาพจริงของข้อมูลโดยไม่มีปัจจัยภายนอกเข้ามาเกี่ยวข้องสามารถเห็นทิศทางการเปลี่ยนแปลงของข้อมูลในช่วงเวลา อีกทั้งยังลดความเอนเอียงอันเนื่องมาจากความรู้หรือเจตคติของผู้วิจัย แต่ทั้งนี้การศึกษาตามยาวก็มีข้อเสียคือหากทำการศึกษาในช่วงระยะเวลาที่ยืดเยื้อมากเกินไปอาจทำให้การแบ่งส่วนของข้อมูลทำได้ยาก เนื่องจากมีอิทธิพลภายในจากที่ข้อมูลที่เก็บจากในช่วงเวลาหนึ่งส่งผลทบต่อข้อมูลที่เก็บจากช่วงเวลาอื่น อาจมีข้อมูลสูญหายอันเนื่องมาจากการเก็บข้อมูล หรือ ผลการวิเคราะห์ข้อมูลอาจไม่ถูกต้องหากเลือกใช้วิธีการวิเคราะห์ทางสถิติที่ไม่เหมาะสม โดยส่วนใหญ่แล้วการศึกษาตามยาวมักมีประโยชน์ในด้านการประเมินความเสี่ยงของโรคหรืองานที่ต้องการศึกษาผลลัพธ์ของทรีตเมนท์ในช่วงระยะเวลาที่แตกต่างกัน

ในทางกลับกันข้อมูลตามขวาง (cross-sectional data) จะเป็นการศึกษาตัวแปรหลายตัวแปรพร้อมกันแต่อาจจะไม่ได้เป็นการศึกษาในช่วงระยะเวลาที่ยาวนานนักก็ได้ดังนั้นการศึกษาตามขวางจะให้ข้อมูลที่มีผลกระทบในด้านความสัมพันธ์ระหว่างตัวแปรที่น้อยกว่า อย่างไรก็ตามแม้ว่าการศึกษาตามขวางจะสามารถทำในช่วงระยะเวลาที่สั้นกว่าแต่การศึกษาประเภทนี้อาจจะต้องทำการพิจารณาการศึกษาตามยาวมาเบื้องต้นก่อน

การวิเคราะห์ข้อมูลตามคาบเวลานั้นสามารถดำเนินการด้วยการวิเคราะห์ความแปรปรวน การวิเคราะห์ความแปรปรวนพหุคูณ หรือ การวิเคราะห์การถดถอยได้เช่นกันหากข้อมูลนั้นตรงตามข้อสมมติเบื้องต้นทางสถิติ(Zaiontz, 2014) บ่อยครั้งที่นักวิจัยเก็บข้อมูลตามคาบเวลาแต่เมื่อทำการวิเคราะห์ข้อมูลกลับใช้สมมติฐานสำหรับข้อมูลตามขวางซึ่งเป็นการเลือกใช้วิธีการทางสถิติที่ไม่ถูกต้องจะทำให้การวิเคราะห์และตีความข้อมูลนั้นเกิดความผิดพลาดและยังจะเป็นการเพิ่มความผิดพลาดแบบที่ 2 ทางสถิติอีกด้วย (Liu, Cripe และ Kim, 2010)

โดยส่วนใหญ่แล้วข้อมูลตามคาบเวลามักพบในงานวิจัยทางด้านสาธารณสุข เนื่องจากส่วนใหญ่จะมีการบันทึกข้อมูลอยู่แล้วในช่วงเวลาที่ยาวนาน ทั้งนี้การเก็บข้อมูลในช่วงเวลาที่ยาวนานนั้นเป็นสาเหตุหนึ่งที่ทำให้เกิดค่าข้อมูลสูญหายที่ไม่สามารถแก้ไขโดยการเก็บใหม่ได้ ตัวอย่างของข้อมูลตามคาบเวลาเช่นการศึกษาโรคหัวใจขาดเลือดในข้อมูลชุด The Framingham Heart Study ในปี 1948 ซึ่งข้อมูลชุดนี้ทำการศึกษาตัวอย่างขนาด 5209 ตัวอย่างจากเมือง Framingham ในประเทศสหรัฐอเมริกาจากผู้ที่มีอายุระหว่าง 30-62 ปี โดยทำการเก็บข้อมูลตัวอย่างเป็นระยะเวลารวมถึง 20 ปี เก็บตัวแปรที่คาดว่าจะมีอิทธิพลต่อการเป็นโรคหัวใจขาดเลือดได้แก่ อายุ น้ำหนัก การสูบบุหรี่ ความดันเลือด ระดับคลอเรสเตอรอลในเลือด และการออกกำลังกาย ซึ่งการศึกษานี้มีประโยชน์มากในด้านของการบอกถึงตัวแปรที่ส่งผลต่อความเสี่ยงต่อการเกิดโรคหัวใจขาดเลือด และ เพิ่มวิธีการรักษาในด้านของปัจจัยที่เกี่ยวข้องด้วย (Caruana, Roman, Hernández, & Solli, 2015)

1.2 ข้อมูลแบบวัดซ้ำ (Repeated Measure data)

ในแผนแบบการทดลองแบบวัดซ้ำมีลักษณะคล้ายกับแผนแบบการวิเคราะห์ความแปรปรวนทางเดียวแต่แตกต่างกันที่สำหรับแผนแบบการทดลองแบบวัดซ้ำหน่วยทดลองในแต่ละกลุ่มจะไม่เป็นอิสระกัน ซึ่งนั่นก็คือเป็นส่วนขยายของ t-test สำหรับข้อมูลที่ไม่เป็นอิสระกัน และตัวอย่างในแต่ละกลุ่มอาจมีลักษณะคล้ายกัน อันเนื่องมาจากตัวอย่างที่ถูกเลือกมีความคล้ายกัน ข้อมูลแบบวัดซ้ำจะเป็นข้อมูลที่เก็บ ตัวแปรแต่ละกลุ่มจากหน่วยตัวอย่างเดียวกันแต่ต่างกันที่ช่วงเวลาที่เก็บซ้ำซึ่งจะเรียกว่าเป็นการวัดซ้ำในหน่วยทดลอง (With-in Subject Repeated Measure) นั่นเอง ในแผนแบบ

การทดลองแบบวัดซ้ำหากมีการวัดซ้ำหลายครั้งจะเรียกว่าเป็นข้อมูลตามคาบเวลาและใช้วิธีการทดสอบในส่วนของการศึกษาตามคาบเวลาแทน (Shuttleworth, 2009) ในแผนแบบการทดลองแบบวัดซ้ำจะเรียกตัวแปรอิสระว่าเป็น โดยข้อมูลแบบวัดซ้ำที่ใช้ในการวิเคราะห์ความแปรปรวนแบบวัดซ้ำมีลักษณะดังแสดงในตารางที่ 1

ตารางที่ 1 รูปแบบของข้อมูลแบบวัดซ้ำ

ตัวอย่าง	เวลา			
	T_1	T_2	...	T_k
S_1	y_{11}	y_{12}	...	y_{1k}
S_2	y_{21}	y_{22}	...	y_{2k}
⋮	⋮	⋮	...	⋮
S_n	Y_{n1}	y_{n2}	...	y_{nk}

เมื่อ y_{ij} คือ ค่าสังเกตที่เก็บมาจากตัวอย่างที่ i ; i = 1, 2, ... , n ในเวลาที่ j ; j = 1, 2, ... , k

ตารางข้างต้นแสดงตัวอย่างข้อมูลแบบวัดซ้ำที่ศึกษาตัวอย่างจำนวน n หน่วยและวัดซ้ำเป็นเวลา k ครั้งจะเห็นว่าในข้อมูลแบบวัดซ้ำจะเรียกแทนทรีตเมนต์ด้วยเวลา และจะเห็นว่าในแต่ละช่วงเวลาจะวัดค่าจากหน่วยตัวอย่างเดียวกันซ้ำกันจึงเรียกว่าเป็นข้อมูลแบบวัดซ้ำ

ข้อตกลงเบื้องต้นสำหรับการวิเคราะห์ความแปรปรวนแบบวัดซ้ำคือหน่วยตัวอย่างแต่ละหน่วยเป็นอิสระกันและถูกสุ่มมาจากประชากรที่มีการแจกแจงปรกติ และนอกจากนี้มีข้อสมมติเรื่อง Sphericityนั่นคือความแปรปรวนของประชากรในแต่ละช่วงเวลาเท่ากันซึ่งคล้ายกับข้อสมมติเรื่องความแปรปรวนของประชากรเท่ากัน ในการวิเคราะห์ความแปรปรวนซึ่งสามารถทดสอบด้วย Mauchly's test แต่ถ้าหากสมมติฐานของ sphericity ถูกปฏิเสธนั่นคือความแปรปรวนของประชากรในแต่ละช่วงเวลาไม่เท่ากันจะสามารถปรับแก้ค่าของผลลัพธ์ด้วยวิธีการปรับแก้ของ Huynh-Feldt หรือ วิธีการปรับแก้ของGreenhouse-Geisser ได้ ("A comprehensive guide to repeated measures ANOVA test," 2019)

ในการวิเคราะห์ความแปรปรวนแบบวัดซ้ำภายในหน่วยทดลองจะทำเพื่อทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากรหลายกลุ่มที่มีความสัมพันธ์กันดังนั้นสมมติฐานว่างของการทดสอบ (H_0) คือค่าเฉลี่ยของประชากรที่มีความสัมพันธ์กันเท่ากันทุกกลุ่ม จะเขียนได้ดังนี้

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

เมื่อ μ คือค่าเฉลี่ยของประชากร และ k คือจำนวนของกลุ่มที่มีความสัมพันธ์กัน ส่วนสมมติฐานแย้งของการทดสอบ (H_1) คือมีค่าเฉลี่ยของประชากรที่มีความสัมพันธ์กันอย่างน้อยหนึ่งกลุ่มแตกต่างจากกลุ่มอื่น

$$H_1 : \mu_i \neq \mu_j ; i, j = 1, 2, \cdots, k \exists i \neq j$$

ข้อดีของการเลือกใช้แผนแบบการทดลองแบบวัดซ้ำที่สำคัญคือจะลดความคลาดเคลื่อนลงจากแผนแบบการวิเคราะห์ความแปรปรวนทางเดียวได้ เนื่องจากในแผนแบบการวิเคราะห์ความแปรปรวนทางเดียวนั้นจะแบ่งแหล่งของความผันแปรซึ่งคำนวณค่าในรูปของค่าผลบวกกำลังสองเป็นแหล่งความผันแปรระหว่างกลุ่ม (Between-groups variability : SSB) ซึ่งเป็นความผันแปรอันเนื่องมาจากเวลาหรือเงื่อนไข (Times variability : SStime) และแหล่งความผันแปรภายในกลุ่ม (Within-groups variability : SSW)โดยจะใช้แหล่งความผันแปรภายในกลุ่มเป็นความคลาดเคลื่อน (Error variability : SSE) แต่ในแผนแบบการทดลองแบบวัดซ้ำนั้นแหล่งความผันแปรภายในกลุ่มจะถูกแบ่งออกเป็นสองส่วน คือ ความคลาดเคลื่อนและความผันแปรอันเนื่องมาจากหน่วยตัวอย่าง (Partitioning of subject variability : SSsubjects) ซึ่งจะทำให้ส่วนของความคลาดเคลื่อนลดลง (LUND & LUND, 2018b) ในแผนแบบการทดลองแบบวัดซ้ำการคำนวณค่าของตัวสถิติทดสอบเอฟสามารถคำนวณได้จากการหาสัดส่วนระหว่างค่าเฉลี่ยของความผันแปรระหว่างกลุ่ม (Mean sum of squares for between-groups : MSB) ซึ่งในที่นี้คือค่าเฉลี่ยของความผันแปรอันเนื่องมาจากความแตกต่างของคาบเวลาหรือเงื่อนไข (Mean sum of squares times : MStime) กับค่าเฉลี่ยของความผันแปรภายในกลุ่ม (Mean sum of squares for within-groups : MSW) หรือในที่นี้จะเรียกว่าเป็นค่าเฉลี่ยของความคลาดเคลื่อน (Mean sum of squares error : MSE) นั่นคือ

$$F = \frac{MStime}{MSE}$$

การสรุปผลในแผนแบบการทดลองแบบวัดซ้ำจะสามารถปฏิเสธสมมติฐานว่างได้หากค่าสถิติทดสอบเอฟที่คำนวณได้มีค่ามากกว่าค่าวิกฤต $f_{\alpha, k-1, (k-1)(n-1)}$ ซึ่งได้จากการเปิดตารางการแจกแจง F

การคำนวณค่าผลบวกกำลังสองและกำลังสองเฉลี่ยสามารถคำนวณได้ดังนี้

$$SStime = SSB = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y} \right)^2$$

$$SSsubjects = k \cdot \sum_{i=1}^{n} \left(\bar{y}_i - \bar{y} \right)^2$$

$$SSW = \sum_{j=1}^{k} \left(y_{ij} - \bar{y}_j \right)^2$$

$$SSW = SSsubjects + SSE$$

$$SST = \sum_{i=1}^{n} \sum_{j=1}^{k} \left(y_{ij} - \bar{y} \right)^2 = SStime + SSsubjects + SSE$$

$$MStime = \frac{SStime}{k-1}$$

$$MSE = \frac{SSE}{(n-1)(k-1)}$$

โดยทั่วไปแล้วการรายงานผลการวิเคราะห์ความแปรปรวนแบบวัดซ้ำมักจะเขียนในรูปของตารางวิเคราะห์ความแปรปรวนในบางครั้งอาจไม่ต้องแสดงความผันแปรของหน่วยตัวอย่างก็ได้ โดยจะแสดงเฉพาะความผันแปรจากช่วงเวลาที่เก็บและความคลาดเคลื่อนที่นำมาใช้คำนวณค่าสถิติทดสอบเอฟเท่านั้น ซึ่งสามารถเขียนได้ดังนี้

ตารางการวิเคราะห์ความแปรปรวนสำหรับแผนแบบการทดลองแบบวัดซ้ำ

ตารางที่ 2 ตารางวิเคราะห์ความแปรปรวน

แหล่งของความผันแปร	ผลบวกกำลังสอง	องศาอิสระ	กำลังสองเฉลี่ย	F
เวลา	SStime	k-1	MStime	MStime/ MSE
หน่วยตัวอย่าง	SSsubjects	n-1	MSsubjects	MSsubjects/ MSE
ความคลาดเคลื่อน	SSE	(k-1)(n-1)	MSE	
รวม	SST	nk-1		

(Field, 2000)

<u>ตัวอย่างที่ 1</u> Dongen, Olofson, Dinges, และ Maislin (2004) ได้ศึกษาเกี่ยวกับผลกระทบของคาเฟอีนต่อผลการนอนหลับโดยวัดค่า Psychomotor vigilance performance laps วัดซ้ำในช่วงเวลา 4 วันเก็บข้อมูลได้ดังแสดงในตารางที่ 3

ตารางที่ 3 ข้อมูลของตัวอย่างที่ 1 ข้อมูลแบบวัดซ้ำ

หน่วยทดลอง	ก่อนทดลอง	วันที่ 1	วันที่ 2	วันที่ 3	$\bar{y}_i$
1	0.4	2.0	23	31.3	14.18
2	1	5.7	19.1	21.9	11.93
3	0.6	7.4	9.9	25.9	10.95
4	4.7	11.3	18	11.7	11.43
5	2.6	5.9	13.6	23.7	11.45
6	1.6	4.0	19.9	24.4	12.48
7	3.3	5.7	11.4	15.4	8.95
8	4.4	4.7	19.6	13.3	10.50
9	0.0	10.1	20.3	12.4	10.70
10	2.6	10.9	8	13.7	8.80
11	0.3	3.3	9.7	13.6	6.73
12	3.0	9.3	6.7	13.4	8.10
13	2.7	3.9	12	12	7.65
$\bar{y}_j$	2.09	6.48	14.71	17.90	$\bar{y}$ =10.29

จากตัวอย่างข้างต้นสามารถคำนวณค่าผลบวกกำลังสองและกำลังสองเฉลี่ยได้ดังนี้

$$SStime = SSB = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2$$

$$= 13 \times \left[(2.09 - 10.29)^2 + \left(6.48 - 10.29\right)^2 + \left(14.71 - 10.29\right)^2 + \left(17.90 - 10.29\right)^2 \right]$$

$$= 2069.21$$

$$SSsubjects = k \cdot \sum_{i=1}^{n} \left(\bar{y}_i - \bar{y}\right)^2$$

$$= 4 \times [\left(14.18 - 10.29\right)^2 + \left(11.93 - 10.29\right)^2 + \ldots + \left(7.65 - 10.29\right)^2]$$

$$= 217.25$$

$$SSW = \sum_{i=1}^{n} \sum_{j=1}^{k} \left(y_{ij} - \bar{y}_j\right)^2$$

$$= [(0.40-2.09)^2 + (1.00-2.09)^2 + \ldots + (2.70-2.09)^2] +$$
$$[(2.00-6.48)^2 + (5.70-6.48)^2 + \ldots + (3.90-6.48)^2] +$$
$$[(23.00-14.71)^2 + (19.10-14.71)^2 + \ldots + (12.00-14.71)^2] +$$
$$[(31.30-17.90)^2 + (31.30-17.90)^2 + \ldots + (12.00-17.90)^2]$$
$$= 1022.40$$

$$SSW = SSsubjects + SSE$$

$$SSE = SSW - SSsubjects$$
$$= 1022.40 - 217.25$$
$$= 805.15$$

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(y_{ij} - \bar{y}\right)^2 = SStime + SSsubjects + SSE$$
$$= 2069.21 + 217.25 + 805.15$$
$$= 3019.61$$

$$F = \frac{MStime}{MSE} = \frac{SStime/k-1}{SSE/(k-1)(n-1)}$$
$$= \frac{2069.21/4-1}{805.15/(4-1)(13-1)}$$
$$= 689.74$$

จากการคำนวณข้างต้นสามารถสร้างตารางวิเคราะห์ความแปรปรวนได้ดังแสดงในตารางที่ 4

ตารางที่ 4 ตารางวิเคราะห์ความแปรปรวนสำหรับตัวอย่างที่ 1

แหล่งของความผันแปร	ผลบวกกำลังสอง	องศาอิสระ	กำลังสองเฉลี่ย	F
เวลา	2069.2	3	689.74	30.84
หน่วยตัวอย่าง	217.3	12	18.10	
ความคลาดเคลื่อน	805.2	36	22.37	
รวม	3091.6	51		

จากค่า F ในตารางความแปรปรวนข้างต้นพบว่าค่าของสถิติทดสอบ F = 30.84 > $f_{0.05,3,36}$ = 2.87 ดังนั้นจึงสามารถปฏิเสธสมมติฐานว่างได้กล่าวคือ ค่า Psychomotor vigilance performance laps หลังจากได้รับคาเฟอีนแล้ว มีอย่างน้อย 1 วันที่มีค่าต่างจากวันอื่นที่ระดับนัยสำคัญ 0.05

และจากตารางข้างต้นจะเห็นว่ามีค่าความคลาดเคลื่อนกำลังสอง(SSE) = 805.2 ซึ่งค่าความคลาดเคลื่อนกำลังสองนี้สามารถคำนวณได้จาก SSW=SSsubjects+SSE ซึ่งค่า SSW นี้จะมีค่าเท่ากับค่าความคลาดเคลื่อนกำลังสองในการวิเคราะห์ความแปรปรวนทางเดียว หมายความว่าหากทำการวิเคราะห์ความแปรปรวนทางเดียวจะมีค่าความคลาดเคลื่อนกำลังสองมากถึง 1022.40 นั่นคือการเลือกใช้แผนแบบการทดลองแบบวัดซ้ำสามารถลดค่าความคลาดเคลื่อนกำลังสองได้ถึง 21.25%

1.3 ชุดข้อมูลจริงที่ใช้ในงานวิจัย

ในการศึกษาครั้งนี้จะประมาณค่าข้อมูลสูญหายโดยสุ่มจากข้อมูลจริง โดยพิจารณาข้อมูลจริง 3 ชุดได้แก่

1.3.1 ข้อมูลชุด Drug Effect (Winer, 1962)

ข้อมูลชุด Drug Effect ถูกเก็บมาเพื่อศึกษาระยะเวลาในการเกิดปฏิกิริยาของยาโดยวัดในยา 4 ชนิด (k = 4) ซึ่งวัดค่าจากตัวอย่างสุ่มขนาด 5 (n = 5) โดยลำดับของยาละชนิด ที่ให้ตัวอย่างแต่ละหน่วยเป็นไปอย่างสุ่ม และมีการเว้นระยะเวลาเพียงพอก่อนที่จะให้ยาตัวใหม่เพื่อป้องกันผลกระทบจากฤทธิ์ยาที่ให้ในลำดับก่อนหน้า

โดยตัวแปรตามคือค่าเฉลี่ยของระยะเวลาในการเกิดปฏิกิริยาของยา และตัวแปรอิสระคือชนิดของยา 4 ชนิด (k=4) ประกอบด้วย ยาชนิดที่ 1, ยาชนิดที่ 2, ยาชนิดที่ 3 และยาชนิดที่ 4

ตารางที่ 5 ข้อมูลชุด Drug Effect

ตัวอย่าง	ยาชนิดที่ 1	ยาชนิดที่ 2	ยาชนิดที่ 3	ยาชนิดที่ 4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

1.3.2 ข้อมูลชุด Skydive (Singley, Hale, & Russell, 2012)

ข้อมูลชุด Sky Drive เป็นข้อมูลจากงานวิจัยเรื่องอัตราการเต้นของหัวใจของนักบินตั้งแต่ก่อนเริ่มบินจนนำเครื่องลงจอดจากนักบินทั้งหมด 11 คน (n = 11) ซึ่งตัวอย่างเป็นนักบินเพศชาย 8 คน และนักบินเพศหญิง 3 คน มีช่วงอายุนะหว่าง 18 – 40 ปี โดยการวัดอัตราการเต้นของหัวใจวัดโดยเครื่อง Polar F6 heart rate monitor ซึ่งในแต่ละครั้งจะทำการวัดเป็นเวลา 1 นาที

โดยตัวแปรตามคืออัตราการเต้นของหัวใจของนักบิน (ครั้ง / นาที) และตัวแปรอิสระคือช่วงเวลาที่ทำการบิน 5 ช่วงเวลา (k=5) ประกอบด้วย

1. ช่วงก่อนขึ้นบิน
2. ช่วงขณะกำลังขึ้นบิน
3. เมื่อเครื่องบินเหนือพื้นดิน 1524 เมตร
4. เมื่อเครื่องบินเหนือพื้นดิน 3048 เมตร
5. เมื่อเครื่องบินกำลังลงจอด

ตารางที่ 6 ข้อมูลชุด Skydive

ตัวอย่าง	ก่อนบิน	ขึ้นบิน	1524 เมตร	3048 เมตร	ลงจอด
1	73.78	101.39	100.13	125.48	91.47
2	79.6	85.85	93.73	132.82	89.23
3	81.37	101.71	86.31	113.04	93.32
4	85.46	87.52	87.27	116.4	89.59
5	85.03	84.51	102.81	125.73	77.07
6	67.81	90.98	92.15	123.28	90.56
7	64.79	87.8	109.77	112.51	79.37
8	84.82	85.92	110.86	121.3	84.2
9	78.31	104.71	101.32	123.94	95.09
10	68.13	92.3	97.45	126.41	71.9
11	76.91	92.27	98.18	122.09	86.18

1.3.3 ข้อมูลชุด Fecal Fat(Vittinghoff, Glidden, Shiboski, & McCulloch, 2012)

ข้อมูลชุดนี้ศึกษาความผิดปรกติของการดูดซึมอาหารในลำไส้ซึ่งอาจเกิดมาจากเอนไซม์ย่อยอาหาร โดยสามารถวัดความผิดปรกติของการดูดซึมอาหารได้จากไขมันส่วนเกินในอุจจาระ ต้องการศึกษาว่ารูปแบบของอาหารเสริม 4 ประเภท (k = 4) มีผลต่อความแตกต่างของไขมันส่วนเกินในอุจจาระหรือไม่ โดยจะเก็บข้อมูลจากตัวอย่างทั้งหมด 6 คน (n = 6)

โดยตัวแปรตามคือปริมาณไขมันส่วนเกินในอุจจาระ (กรัม / วัน) ตัวแปรอิสระคือรูปแบบของอาหารเสริม 4 ประเภท (k = 4) ประกอบด้วย

1. ยาหลอก (placebo)
2. ยาอัดเม็ด (tablet)
3. แคปซูลไม่เคลือบ (capsule)
4. แคปซูลเคลือบ (coated)

ตารางที่ 7 ข้อมูลชุด Fecal Fat

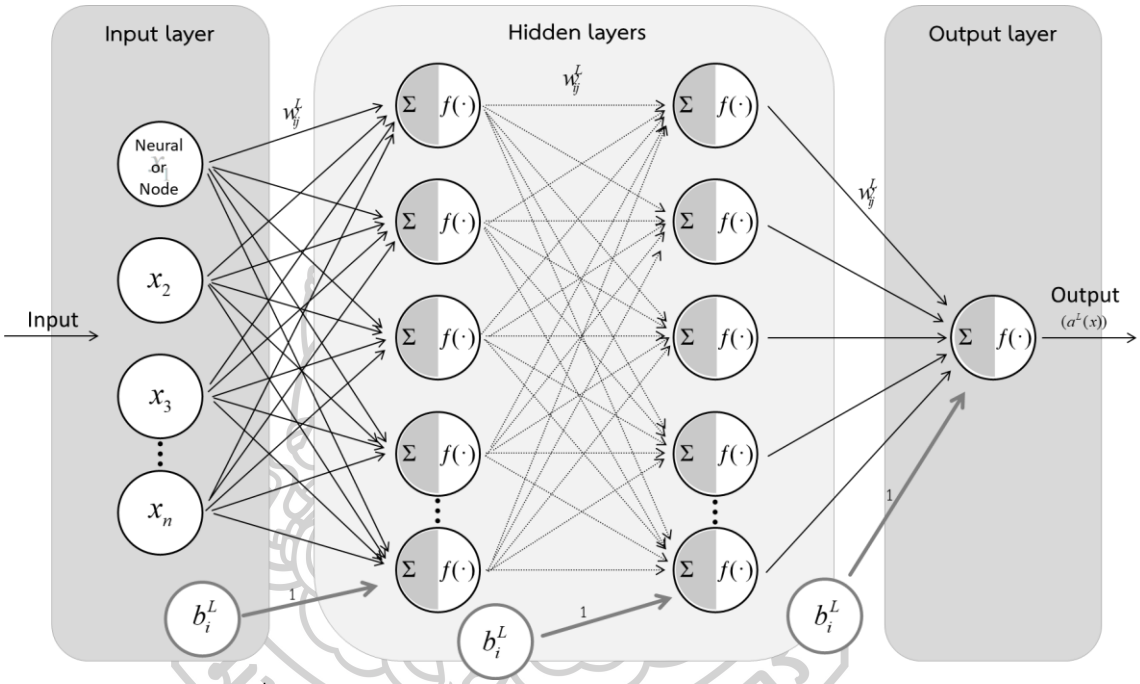
ตัวอย่าง	ยาหลอก	ยาอัดเม็ด	แคปซูลไม่เคลือบ	แคปซูลเคลือบ
1	44.5	7.3	3.4	12.4
2	33	21	23.1	25.4
3	19.1	5	11.8	22
4	9.4	4.6	4.6	5.8
5	71.3	23.3	25.6	68.2
6	51.2	38	36	52.6

1.4 โครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทเทียมเป็นเครื่องมือทางคอมพิวเตอร์ที่มีประโยชน์หลากหลายเช่นการตรวจจับหรืออ่านค่าข้อมูลต่าง ๆ เช่น ลายมือ ลายนิ้วมือ เสียงพูด และยังสามารถใช้ในการประมาณค่า และพยากรณ์ได้อีกด้วย โดยวิธีการโครงข่ายประสาทเทียมมีแนวคิดมาจากการทำงานของระบบประสาทของสิ่งมีชีวิตแต่ระบบการเรียนรู้ของระบบประสาทและกระบวนการทางคอมพิวเตอร์มีความแตกต่างกัน เช่นในสิ่งมีชีวิตการตัดสินใจว่า “9” คือเลขเก้านั้นมนุษย์จะตัดสินใจจากรูปร่างโดยเลขเก้ามีส่วนประกอบคือ มีวงกลมอยู่ด้านบนและมีหางต่อมาทางด้านขวาของวงกลม แต่ด้วยการใช้การตัดสินใจเช่นนี้เมื่อมาประยุกต์ใช้กับกระบวนการทางคอมพิวเตอร์อาจเกิดข้อผิดพลาดได้หากใช้ตรวจจับลายมือของบุคคลที่มีวิธีการเขียนที่ต่าง หรือซับซ้อนกว่าผู้อื่น สำหรับแนวคิดของโครงข่ายประสาทเทียมมีวิธีการในการตรวจจับคือนำเข้าข้อมูลเลขเก้าที่เขียนจากลายมือจำนวนหลาย ๆ ลายมือจากนั้นใช้การเรียนรู้จากตัวอย่างเพื่อตรวจจับว่า “9” คือเลขเก้า ซึ่งการเพิ่มขนาดตัวอย่างก็จะทำให้โครงข่ายมีการเรียนรู้เพิ่มขึ้นและเพิ่มความแม่นยำในการตรวจจับข้อมูลได้มากขึ้นอีกด้วย (Nielson, 2019)

Rosenblatt (1957) ได้เสนอกระบวนการที่เรียกว่า Perceptron ที่เป็นกระบวนการที่ส่งออกผลลัพธ์จากของค่าน้ำหนัก ซึ่งในปัจจุบันนำมาใช้เป็นตัวแบบในรูปทั่วไปของโครงข่ายประสาทเทียมกันอย่างแพร่หลาย กระบวนการของ Proceptron คือ นำเข้าตัวแปร $x_1, x_2, \cdots, x_n$ เป็นตัวแปรนำเข้าจากนั้นนำเข้าน้ำหนัก $w_1, w_2, \cdots w_n$ ที่เป็นจำนวนจริงเพื่อเป็นตัวบ่งบอกถึงความสำคัญของตัวแปรนำเข้าแต่ละตัวที่มีอิทธิพลต่อผลลัพธ์ ซึ่งผลลัพธ์นั้นจะคำนวณได้ผลรวมถ่วงน้ำหนักของตัวแปร

นำเข้า ซึ่งเขียนได้ดังนี้ $\sum_j w_j x_j$ หรือหากพิจารณา x และ w ในรูปของเวกเตอร์สามารถเขียนผลรวมถ่วงน้ำหนักในรูปของผลคูณเชิงสเกลาร์ได้ดังนี้ $w \cdot x \equiv \sum_j w_j x_j$ แล้วค่าของผลลัพธ์ที่ได้จะขึ้นอยู่กับค่าของผลคูณเชิงสเกลาร์ของค่าของเวกเตอร์ของตัวแปรนำเข้าและเวกเตอร์ของน้ำหนักบวกกับเวกเตอร์ความเอนเอียง $w \cdot x + b$ ใน activation function ที่กำหนด ซึ่งการเลือก activation function ที่เหมาะสมกับลักษณะของผลลัพธ์ที่ต้องการจะส่งผลให้ค่าของผลลัพธ์ถูกต้องมากขึ้นซึ่งจะสามารถเขียนโครงสร้างของโครงข่ายประสาทเทียมเป็นแผนภาพได้ดังนี้



ภาพที่ 1 แผนภาพกระบวนการทำงานของโครงข่ายประสาทเทียม

ในส่วนซ้ายสุดของโครงข่ายเรียกว่า Input layer และจุดหรือนิวรอนภายในจะเรียกว่า Input neurons ในส่วนขวาสุดเรียกว่า Output layer ซึ่งจุดหรือนิวรอนภายในจะเรียกว่า Output neurons ซึ่งอาจกำหนดให้มีค่าเดียวหรือหลายค่าก็ได้ และในส่วนกลางจะเรียกว่า Hidden layers โดยในโครงข่ายจะมีสามารถมี Hidden layers ได้มากกว่า 1 เลเยอร์ โดยหากมี Hidden layers มากกว่า 1 เลเยอร์จะเรียกตัวแบบว่า Multilayer perceptrons (MLPs) โครงข่ายประสาทเทียมจะใช้ผลลัพธ์ของเลเยอร์หนึ่งจะใช้เป็นตัวแปรนำเข้าของเลเยอร์ถัดไปไปเรื่อย โดยผลลัพธ์ในแต่ละเลเยอร์คำนวณจากการคำนวณฟังก์ชันของผลรวมของผลคูณระหว่าง a_j^{l-1} (ตัวแปรนำเข้าในนิวรอนที่ j จากเลเยอร์ที่ (l-1)) กับ w_{ij}^l (น้ำหนักระหว่างนิวรอนที่ i ในเลเยอร์ที่ (l-1) และนิวรอนที่ j ในเลเยอร์ที่ l) บวกกับ b_j^l (ค่าความเอนเอียงในนิวรอนที่ j ในเลเยอร์ที่ l) นั่นคือ

$$a_i^l = f\left(\sum_j w_{ij}^l a_j^{l-1} + b_i^l\right)$$

เมื่อแทน $\sum_j w_{ij}^l a_j^{l-1} + b_i^l$ ด้วย z^l จะได้ $a_i^l = f\left(z^l\right)$

การปรับค่าน้ำหนักนั้นจะพิจารณาจากฟังก์ชั้นความคลาดเคลื่อนกำลังสอง (MSE) ของเวกเตอร์ผลลัพธ์ $y(x)$ และเวกเตอร์ของผลลัพธ์จากโครงข่าย $a^L(x)$ เมื่อผลลัพธ์ของโครงข่ายคำนวณมาจากเวกเตอร์ของตัวแปรนำเข้า (x) เวกเตอร์ของน้ำหนัก (w) และเวกเตอร์ของความเอนเอียง (b) ดังนี้

$$C(w,b) \equiv \frac{1}{2n}\sum_x \left\| y(x) - a^L(x) \right\|^2$$

หรือสามารถเขียนฟังก์ชันความคลาดเคลื่อนของผลลัพธ์แต่ละค่าได้ดังนี้

$$C = \frac{1}{2}\left\| \boldsymbol{y} - \boldsymbol{a^L} \right\|^2 = \frac{1}{2}\sum_i \left(y_i - a_i^l\right)^2$$

เมื่อพิจารณากระบวนการเรียนรู้ของโครงข่ายประสาทเทียมจะเห็นว่าหากเพิ่มผลการเปลี่ยนแปลงเล็ก ๆ ในค่าน้ำหนักหรือค่าความเอนเอียงจะทำให้ผลลัพธ์มีค่าเปลี่ยนแปลงไปเล็กน้อยด้วยเช่นกัน ดังนั้นเราจึงสามารถใช้การเปลี่ยนแปลงของผลลัพธ์มาปรับน้ำหนักได้ ซึ่งการปรับค่าน้ำหนักเป็นการเพิ่มความถูกต้องของการพยากรณ์ผลลัพธ์การคำนวนค่าความคลาดเคลื่อน (δ_i^l) จากค่าของน้ำหนักที่เปลี่ยนแปลงไปเล็กน้อย Δw_{ij}^l และค่าของความเอนเอียงที่เปลี่ยนไปเล็กน้อย Δb_i^l ซึ่งสามารถประมาณค่าของผลลัพธ์ที่เปลี่ยนแปลงไปเล็กน้อยนี้จาก $\partial C/\partial w_{ij}^l$ และ $\partial C/\partial b_i^l$ เมื่อแทนฟังก์ชั้นของน้ำหนักและความเอนเอียงด้วย z^l แล้วดังนั้นจะสามารถคำนวณค่าความคลาดเคลื่อนของนิวรอนที่ i จากเลเยอร์ผลลัพธ์ ได้จาก $\delta_i^L = \frac{\partial C}{\partial a_i^L} f'\left(z_i^L\right)$ หรือสามารถเขียนในรูปของเมทริกซ์ได้ดังนี้

$$\boldsymbol{\delta^L} = \nabla_{\boldsymbol{a}} \boldsymbol{C} \odot f'\left(z_i^L\right)$$

เมื่อ $\odot$ คือHadamard productซึ่ง $\left(\boldsymbol{a} \odot \boldsymbol{b}\right)_{\boldsymbol{j}} = a_j b_j$ และเมื่อแทนค่าการเปลี่ยนแปลงของ C ด้วย $(\boldsymbol{a^L} - \boldsymbol{y})$ จะได้

$$\boldsymbol{\delta}^L = (a^L - y) \odot f'\left(z_i^L\right)$$

ในกระบวนการ Backpropagation การปรับค่าน้ำหนักสามารถทำได้โดยการคำนวณความคลาดเคลื่อนจากค่าของน้ำหนักในเลเยอร์ถัดไปดังนี้

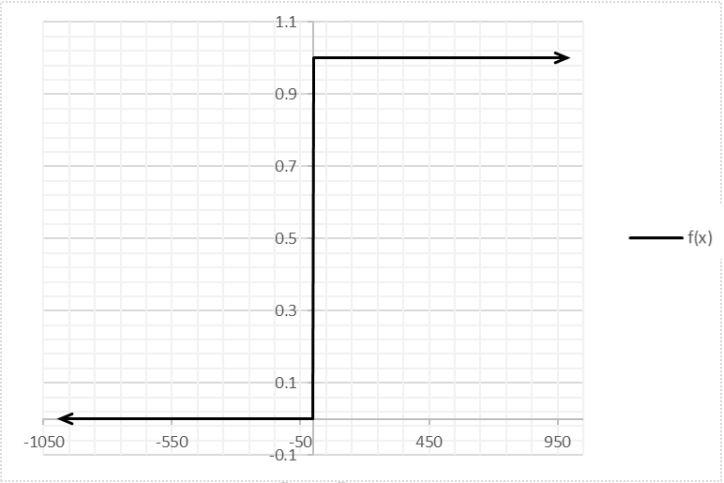
$$\boldsymbol{\delta^L} = ((\boldsymbol{w^{l+1}})^{\boldsymbol{T}}\, \boldsymbol{\delta^{l+1}}) \odot f'\left(z_i^L\right)$$

ส่วนประกอบที่สำคัญอีกอย่างของโครงข่ายประสาทเทียม คือ Activation function เป็นส่วนประกอบหนึ่งที่มีความสำคัญมากในโครงข่ายประสาทเทียมที่ทำให้โครงข่ายประสาทเทียมเป็นกระบวนการที่ไม่เป็นเชิงเส้นโดยพื้นฐานแล้วโครงข่ายประสาทเทียมมีส่วนประกอบคือตัวแปรนำเข้า (x) ค่าน้ำหนัก(w) และส่งออกค่าผลลัพธ์ผ่านฟังก์ชัน $f(x)$ เพื่อนำค่าผลลัพธ์ที่ได้ไปเป็นตัวแปรนำเข้าในเลเยอร์ถัดไป ซึ่งฟังก์ชัน $f(x)$ ในโครงข่ายประสาทเทียมจะเรียกว่าเป็น Activation function หากไม่มี Activation function ผลลัพธ์ของโครงข่ายประสาทเทียมจะเป็นผลลัพธ์จากฟังก์ชันเชิงเส้นธรรมดาซึ่งมีลักษณะเหมือนกับการวิเคราะห์การถดถอยเชิงเส้นแต่ในข้อมูลจริง ข้อมูลอาจไม่ได้มีความสัมพันธ์เชิงเส้นก็ได้ และโดยทั่วไปการคำนวณค่าของผลลัพธ์สามารถมีค่าที่เป็นไปได้ตั้งแต่ลบอนันต์ถึงอนันต์ซึ่งโครงข่ายประสาทเทียมถูกออกแบบมาให้สามารถปรับเปลี่ยนฟังก์ชันในการคำนวณได้ใช้ได้ดีในข้อมูลที่ไม่เป็นเชิงเส้นและสามารถปรับค่าของผลลัพธ์ให้อยู่ในช่วงที่กำหนดได้อีกด้วย ซึ่ง Activation function ที่นิยมใช้มีดังนี้

- ฟังก์ชันขั้นบันได (Step Function)

ฟังก์ชันขั้นบันไดเป็นฟังก์ชันที่แสดงค่าผลลัพธ์ที่เป็นไปได้ 2 ค่า คือค่า 0 และ 1 เหมาะสำหรับใช้ในการจัดกลุ่มในเลเยอร์ผลลัพธ์ แต่ฟังก์ชันขั้นบันไดไม่เหมาะสมที่จะใช้ใน Hidden layer เนื่องจากเป็นฟังก์ชันที่ไม่สามารถหาอนุพันธ์ได้จึงไม่สามารถใช้ในการปรับค่าน้ำหนักได้ ฟังก์ชันและกราฟของฟังก์ชันขั้นบันไดมีลักษณะดังนี้

$$f(x) = \begin{cases} 0 & for \quad x < 0 \\ 1 & for \quad x \geq 0 \end{cases}$$

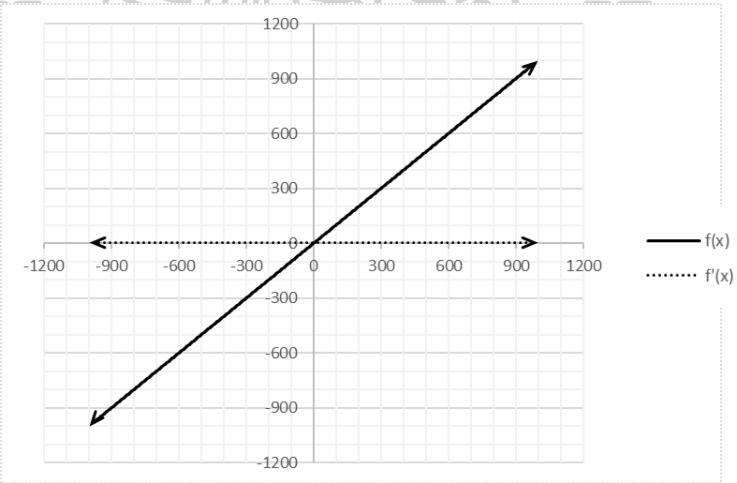


ภาพที่ 2 กราฟของฟังก์ชันขั้นบันได

- ฟังก์ชันเชิงเส้น (Linear Function)

ฟังก์ชันเชิงเส้น หรือ Linear function เป็น Activation function ที่เหมาะสมกับข้อมูลเชิงปริมาณแต่ปัญหาหนึ่งของฟังก์ชันเชิงเส้นคืออนุพันธ์ของฟังก์ชันเป็นค่าคงที่ ซึ่งจะทำให้กระบวนการเรียนรู้หยุดอยู่ที่เดิม และนอกจากนี้ส่งผลให้ให้ผลลัพธ์ในแต่ละเลเยอร์มีค่าเหมือนกัน เลเยอร์ก่อนหน้าจะมีอิทธิพลต่อ เลเยอร์ถัดไปเป็นอย่างมากฟังก์ชันและกราฟของฟังก์ชันเชิงเล้นมีลักษณะดังนี้

$$f(x) = x$$

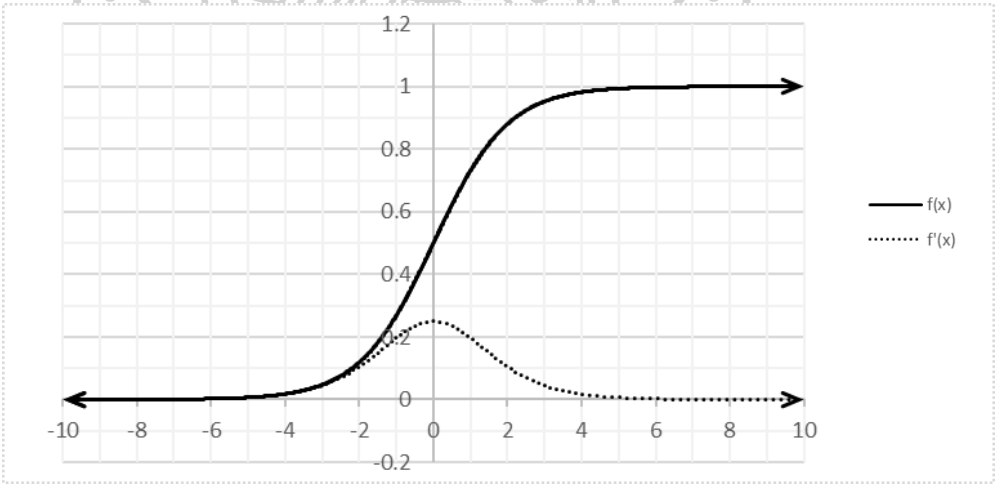


ภาพที่ 3 กราฟของฟังก์ชันเชิงเส้น

- ฟังก์ชันเส้นโค้งซิกมอยด์ (Sigmoid Function)

ฟังก์ชันเส้นโค้งซิกมอยด์มีประโยชน์สำหรับการพยากรณ์ผลลัพธ์ที่เป็นข้อมูลเชิงกลุ่มเนื่องค่าค่าผลลัพธ์ที่เป็นไปได้มีค่าระหว่าง 0 ถึง 1 ซึ่งสามารถจัดกลุ่มของตัวแปรได้จากการเลือกจุดแบ่งของผลลัพธ์ และเนื่องจากข้อมูลในธรรมชาติโดยทั่วไปแล้วจะเป็นข้อมูลไม่เป็นเชิงเส้นฟังก์ชันเส้นโค้งซิกมอยด์ก็เป็นฟังก์ชันที่ไม่เป็นเชิงเส้นเช่นกันฟังก์ชันนี้จึงเหมาะสมกับการวิเคราะห์ข้อมูลส่วนใหญ่ นอกจากนี้ฟังก์ชัน เส้นโค้งซิกมอยด์ ยังเป็นฟังก์ชันที่สามารถหาอนุพันธ์ได้ หมายความว่าจะสามารถคำนวณค่า Learning rate ได้และสารถปรับค่าน้ำหนักในรอบการคำนวณถัดไปได้ฟังก์ชันเส้นโค้งซิกมอยด์เป็นฟังก์ชันหนึ่งที่นิยมใช้อย่างมากสำหรับการจัดกลุ่มแต่ปัญหาของฟังก์ชันเส้นโค้งซิกมอยด์คือเมื่อพิจารณาจารกราฟของฟังก์ชันเส้นโค้งซิกมอยด์จะเห็นว่าค่าผลลัพธ์ ในแกน y เปลี่ยนแปลงไปน้อยมากเมื่อเทียบกับค่าของตัวแปรนำเข้าในแกน x ซึ่งค่าผลลัพธ์จะเปลี่ยนแปลงอย่างรวดเร็วเฉพาะที่ค่าของตัวแปรนำเข้าอยู่ในช่วง -2 ถึง 2 เท่านั้นและอนุพันธ์ของฟังก์ชันเส้นโค้งซิกมอยด์ยังมีค่าน้อยและง่ายต่อการลู่เข้าสู่ศูนย์ นั่นคือค่า Learning rate มีค่าต่ำทำให้กระบวนการเรียนรู้ในรอบถัดไปเกิดขึ้นช้าหรือไม่เกิดเลย และจะส่งผลให้ค่าความคลาดเคลื่อนตกอยู่ที่ค่าต่ำสุดสัมพัทธ์ไม่ใช่ค่าต่ำสุดสัมบูรณ์และทำให้ประสิทธิภาพของการพยากรณ์ด้วยโครงข่ายประสาทเทียมน้อยลงอีกด้วย ฟังก์ชันและกราฟของฟังก์ชันเส้นโค้งซิกมอยด์มีลักษณะดังนี้

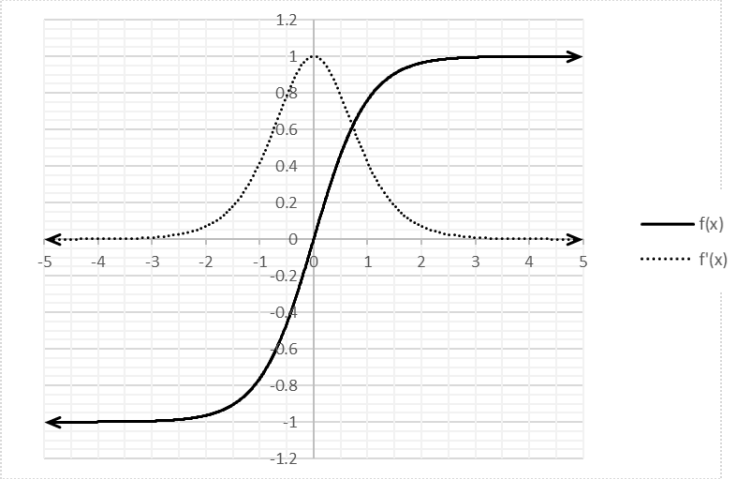
$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$$



ภาพที่ 4กราฟของฟังก์ชันเส้นโค้งซิกมอยด์

- ฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ (Hyperbolic Tangent Function)

ฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ลักษณะคล้ายกับฟังก์ชันเส้นโค้งซิกมอยด์แต่มีค่าที่เป็นไปได้อยู่ในช่วง -1 ถึง 1 และค่าของอนุพันธ์ของฟังก์ชันไฮเพอร์โบลิกแทนเจนต์จะมีความชันมากกว่าฟังก์ชันเส้นโค้งซิกมอยด์หมายความว่าฟังก์ชันนี้มีประสิทธิภาพมากกว่าฟังก์ชันเส้นโค้งซิกมอยด์เนื่องจากสามารถสร้างค่าที่คำนวณได้ในช่วงที่กว้างกว่า และกระบวนการเรียนรู้จะรวดเร็วกว่าด้วย ฟังก์ชันและกราฟของฟังก์ชันไฮเพอร์โบลิกแทนเจนต์มีลักษณะดังนี้

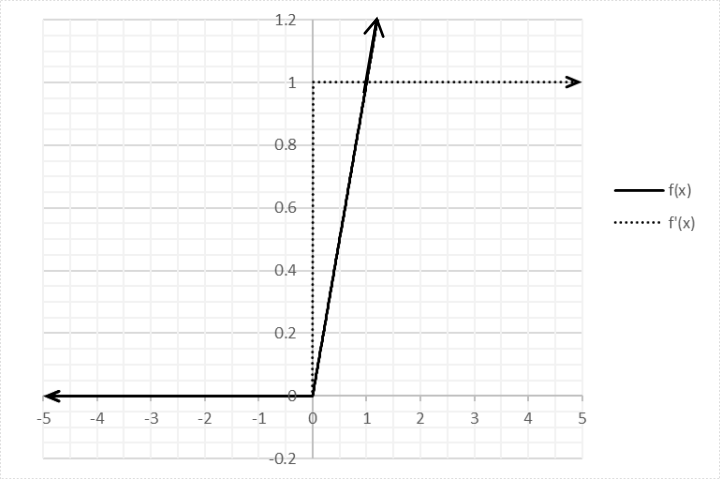


ภาพที่ 5 กราฟของฟังก์ชันไฮเพอร์โบลิกแทนเจนต์

- ReLU (Rectified Linear Unit) Function

ReLU Function ในแกนด้านขวาที่มีค่าเป็นบวกมีลักษณะเหมือนกับฟังก์ชันเชิงเส้น ซึ่งเหมาะสำหรับข้อมูลที่มีลักษณะเป็นเชิงปริมาณที่มีค่าไม่เป็นลบ แต่จะเห็นได้ว่า ReLU Function ไม่ได้เป็นฟังก์ชันเชิงเส้นโดยธรรมชาติ ซึ่งอาจใช้ฟังก์ชันนี้ร่วมกับฟังก์ชันอื่นได้ ค่าที่เป็นไปได้ของ ReLU Function อยู่ในช่วง 0 ถึงอนันต์ ซึ่งการที่อนุพันธ์ของฟังก์ชันนี้มีค่าเป็น 0 ในแกนลบทำให้การรันค่าในโครงข่ายรวดเร็วขึ้น แต่ฟังก์ชันนี้ไม่ได้ผลให้กระบวนการเรียนรู้เร็วขึ้นไปด้วยเนื่องจากในแกนด้านขวายังใช้ฟังก์ชันเชิงเส้นที่มีอนุพันธ์เป็นค่าคงที่อยู่เช่นเดิม ฟังก์ชันและกราฟของ ReLU Function มีลักษณะดังนี้

$$f(x) = \begin{cases} 0 & for \quad x < 0 \\ x & for \quad x \geq 0 \end{cases}$$

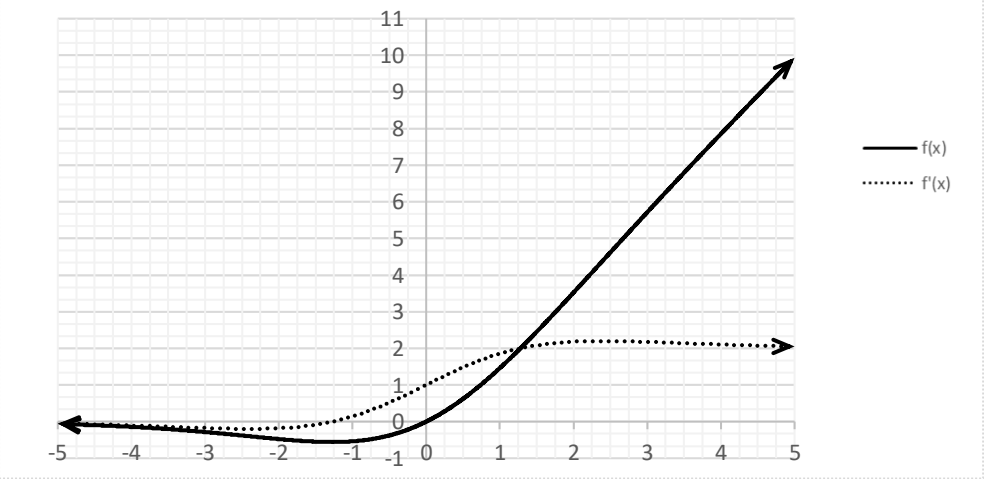


ภาพที่ 6 กราฟของ ReLU Function

- Swish (A Self-Gated) Function

หากกำหนดค่า β = 0 ฟังก์ชัน Swish จะมีลักษณะเป็นฟังก์ชันเชิงเส้นปรกติ แต่ในทางกลับกัน หากเปลี่ยนค่า β เป็นค่ามาก ๆ ฟังก์ชัน Swish จะมีลลักษณธคล้ายกับ ReLU function นั่นคือ มีค่าเป็น 0 หาก x มีค่าน้อยกว่า 0 และเป็นฟังก์ชันเชิงเส้น ($f(x)=2x$) ถ้า x มีค่ามากกว่า 0 อย่างไรก็ตามโดยทั่วไปแล้ว ฟังก์ชัน Swish จะกำหนดค่า β = 1 ฟังก์ชันจะมีลักษณะดังกราฟด้านล่างนี้ ซึ่งจะเห็นว่าในแกนฝั่งขวากราฟมีลักษณะใกล้เคียงกับฟังก์ชันเชิงเส้นอีกทั้งยังสามารถหาอนุพันธ์ได้ทั้งสำหรับค่าของ ตัวแปรนำเข้าที่เป็นบวกและลบ ทำให้สามารถปรับค่าน้ำหนักในกระบวนการในรอบถัดไปได้ ฟังก์ชันและกราฟของฟังก์ชัน Swish มีลักษณะดังนี้

$$f(x)=2x\sigma(\beta x)=\frac{2x}{1+e^{-\beta x}}$$



ภาพที่ 7 กราฟของฟังก์ชัน Swish

การเลือกใช้ Activation function ควรคำนึงลักษณะของข้อมูลที่ศึกษา และควรเลือก Activation function ที่ทำให้กระบวนการเรียนรู้ของโครงข่ายเป็นไปอย่างรวดเร็วที่สุด ตัวอย่างเช่น Sigmoid function เหมาะสมกับข้อมูลแบบแบ่งกลุ่ม ซึ่งหากข้อมูลที่ศึกษาไม่ได้มีความซับซ้อนมาก การเลือกใช้ Sigmoid function เป็น Activation function จะทำได้ง่ายกว่าการเลือกใช้ฟังก์ชัน ReLu และการวิเคราะห์โครงข่ายอาจจะลู่เข้าเร็วกว่าอีกด้วย(Kızrak, 2019) หรือนอกจากนี้หากทราบลักษณะของข้อมูลอยู่แล้วการเลือกใช้ฟังก์ชันอื่น ๆ นอกจากฟังก์ชันข้างต้นก็สามารถทำได้เช่นกัน

การทำงานของโครงข่ายข่ายประสาทเทียมในภาพรวมคือเริ่มจากการนำเข้าตัวแปรนำเข้า จากนั้นคำนวณผลรวมของผลคูณระหว่างค่าของตัวแปรนำเข้ากับค่าน้ำหนัก แล้วนำค่าที่คำนวณได้ไปคำนวณค่าผลลัพธ์ของเลเยอร์อีกครั้งผ่าน Activation function แล้วจะได้ผลลัพธ์ของเลเยอร์เพื่อนำไปใช้เป็นตัวแปรนำเข้าในเลเยอร์ถัดไปจนได้ผลลัพธ์ในเลเยอร์ผลลัพธ์ จากนั้นทำการปรับค่าน้ำหนักแล้วแล้วใช้ค่าน้ำหนักที่ปรับใหม่ในการคำนวณผลลัพธ์ของเลเยอร์ในรอบถัดไป

1.5 การประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS)

การแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย(Mean Substitution) เป็นวิธีการที่ตัดเอาผลกระทบของความแปรปรวนจากตัวอย่างออกซึ่งมีอิทธิพลต่อการทดสอบทางสถิติที่ตามมาและมีการทดสอบแล้วว่าเป็นวิธีการประมาณค่าที่เอนเอียงแต่มีความคงเส้นคงวา(Bingham, Stemmler, Peterson, & Graber, 1998) ซึ่งกระบวนการของการแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย(Mean Substitution) สามารถทำโดยขั้นตอนดังต่อไปนี้

เมื่อกำหนดให้ i คือ index ของค่าสังเกตที่เก็บจากตัวแปรที่ i

j คือ index ของค่าสังเกตที่เก็บในเวลาที่ j

N_j คือจำนวนของตัวแปรที่เก็บในเวลาที่ j

t' คือจำนวนซ้ำของสังเกตที่เก็บจากตัวแปรที่ i ที่ไม่เป็นค่าข้อมูลสูญหาย

1.5.1 หาค่าเฉลี่ยตามขวาง (cross-section mean) โดย $\bar{Y}_{.j} = \frac{1}{N_j}\sum_i y_{ij}$

1.5.2 หาค่าเฉลี่ยของผลต่างระหว่างค่าเฉลี่ยตามขวางกับค่าสังเกตแต่ละค่าโดย

$$\bar{I}_i = \frac{1}{t'}\sum_i (\bar{y}_{.j} - y_{ij})$$

1.5.3 แทนที่ค่าข้อมูลสูญที่ด้วยการแทนที่ด้วยค่าเฉลี่ยที่คำนวณด้วย $\hat{Y}_{ij} = \bar{y}_{.j} - \bar{I}_i$

ตัวอย่างการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย

<u>ตัวอย่างที่ 2</u> ข้อมูลน้ำหนักของผมที่เพิ่มขึ้น (มิลลิกรัม/ตารางเซนติเมตร) หลังจากใช้ยา Minoxidil ในการบำรุงเส้นผมโดยเก็บข้อมูลจาก 5 ช่วงเวลา(Price & Menefee, 1990) ข้อมูลดังแสดงในตารางที่ 8 จากตัวอย่าง ขนาดเท่ากับ 4

ตารางที่ 8 ตัวอย่างที่ 2 ข้อมูลน้ำหนักของผมหลังใช้ยา Minoxidil

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290	146	193	130
16 สัปดาห์	340	206	218	144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

สุ่มค่าข้อมูลสูญหายแบบสุ่มสมบูรณ์ข้อมูลสูญหาย 2 ค่าได้แก่ y_{22} และ y_{33}

ตารางที่ 9 ข้อมุลสูญหายจากตัวอย่างที่ 2

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290		193	130
16 สัปดาห์	340	206		144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

1) คำนวณค่าเฉลี่ยตามขวาง

ตารางที่ 10 ค่าเฉลี่ยตามขวางของตัวอย่างที่ 2 ในตารางที่ 6

เวลา (สัปดาห์)	หน่วยทดลอง				$\overline{y_{.j}}$
	1	2	3	4	
ก่อนทดลอง	216	130	206	106	164.5
8 สัปดาห์	290		193	130	204.33
16 สัปดาห์	340	206		144	230
24 สัปดาห์	275	220	223	150	217
32 สัปดาห์	294	209	226	173	225.5

2) หาค่าเฉลี่ยของผลต่างระหว่างค่าเฉลี่ยตามขวางกับค่าสังเกตแต่ละค่า

$$\bar{I}_2 = \frac{(164.5-130)+(230-206)+(217-220)+(225.5-209)}{4} = 18$$

$$\bar{I}_3 = \frac{(164.5-206)+(204.33-193)+(217-223)+(225.5-226)}{4} = -9.167$$

ประมาณค่าข้อมูลสูญที่ด้วยการแทนที่ด้วยค่าเฉลี่ยที่คำนวณด้วย $\hat{Y}_{ij} = \bar{y}_{.j} - \bar{I}_i$ ดังนั้นจะได้ค่าประมาณข้อมูลสูญหายคือ

$$\hat{y}_{22} = 204.33 - 18 = 186.33$$

$$\hat{y}_{33} = 230 - (-9.167) = 239.167$$

ดังนั้นจะได้ชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS) ดังแสดงในตารางที่ 11

ตารางที่ 11 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290	**186.33**	193	130
16 สัปดาห์	340	206	**239.167**	144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

1.6 การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean

วิธี CopyMean คือกระบวนการประมาณค่าข้อมูลสูญหายที่มีกระบวนการทำ 2 ขั้นตอนที่รวมการประมาณค่าข้อมูลสูญหายจากวิธีการของการศึกษาตามคาบเวลา และการศึกษาตามขวางเข้าด้วยกันได้แก่

1) การใช้การประมาณค่าข้อมูลสูญหายด้วยวิธีการการศึกษาตามคาบเวลา เป็นการประมาณค่าข้อมูลสูญหายโดยใช้ข้อมูลจากตัวแปรเดียวกันเท่านั้นมาประมาณค่าข้อมูลสูญหายในตัวแปรนั้น ๆ ซึ่งการประมาณค่าข้อมูลสูญหายโดยวิธีนี้ค่าที่ประมาณได้จะเป็นอิสระจากตัวแปรอื่น ๆ คือใช้ข้อมูลของ $y_{i.} =(y_{i1}, y_{i2}, ..., y_{it})$ ในการประมาณค่าข้อมูลสูญหายเท่านั้น เช่น ค่าเฉลี่ยตามคาบเวลา ค่ามัธยฐานตามคาบเวลา ค่าสุ่มตามคาบเวลา วิธีการแทนที่ด้วยการวัดความรู้ครั้งสุดท้าย และวิธีการประมาณค่าในช่วงเชิงเส้น เป็นต้น

2) ใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตามขวางซึ่งเป็นการประมาณค่าข้อมูลสูญหายโดยใช้ข้อมูลจากค่าสังเกตที่เก็บในเวลาเดียวกันเท่านั้นมาประมาณค่าข้อมูลสูญหายในเวลานั้น ซึ่งการประมาณค่าข้อมูลสูญหายโดยวิธีนี้ค่าที่ประมาณได้จะเป็นอิสระจากค่าสังเกตในช่วงเวลาอื่น โดยใช้ข้อมูลของ $y_{.j} =(y_{1j}, y_{2j}, ..., y_{nj})$ ในการประมาณค่าข้อมูลสูญหายเท่านั้น ซึ่งมีวิธีการหลังจากที่แทนค่าข้อมูลสูญหายที่ประมาณค่าด้วยวิธีการของการศึกษาตามคาบเวลาแล้ว จากนั้นคำนวณค่าความผันแปรเฉลี่ย (Average Variation : AV) ที่เวลาที่ j โดยคำนวณจากการผลต่างของค่าเฉลี่ยของค่าสังเกตในเวลาที่ j ที่ตัดค่าข้อมูลสูญหายออกแล้ว กับค่าเฉลี่ยของ ค่าสังเกตในเวลาที่ j ที่ตัดค่าที่แทนที่ค่าสูญหายด้วยการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตาม

คาบเวลาแล้ว $\left(AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}\right)$ แล้วนำค่า ความผันแปรเฉลี่ยที่เวลาที่ j ไปบวกกับค่าประมาณจากการประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายตามคาบเวลาเดิม $\left(\hat{y}_{ij} = y_{ij}^{IM} + AV_j\right)$ ค่าประมาณค่าข้อมูลสูญหายที่คำนวณได้ $\hat{y}_{ij}$ จะเป็นค่าประมาณข้อมูลสูญหายจากวิธี CopyMean(Genolini, Lacombe, cochard, & Subtil, 2016)

1.6.1 วิธี CopyMean Trajectory

การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean Trajectory มีขั้นตอนดังนี้

1) คำนวณค่าข้อมูลสูญหายจากวิธีการ Trajectory Mean โดยการแทนที่ค่าข้อมูลสูญหายด้วยค่าเฉลี่ยตามคาบเวลา($\overline{y_{i.}}$)
2) คำนวณค่าความผันแปรเฉลี่ยจากการนำค่าเฉลี่ยของข้อมูลที่มีค่าข้อมูลสูญหายลบด้วยค่าเฉลี่ยของชุดข้อมูลที่แทนที่ค่าข้อมูลสูญหายด้วยการประมาณค่าด้วยวิธีการ Trajectory Mean แล้ว $\left(AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}\right)$
3) คำนวณค่าข้อมูลสูญหายด้วยการนำข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ วิธีการ Trajectory Mean ไปบวกกับค่าความผันแปรเฉลี่ย $\left(\hat{y}_{ij} = y_{ij}^{IM} + AV_j\right)$

<u>ตัวอย่างที่ 3</u> การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean Trajectory โดยใช้ตัวอย่างจากชุดข้อมูลน้ำหนักผมที่สุ่มค่าข้อมูลสูญหายแล้วจากตัวอย่างที่ 2 ในตารางที่ 6 โดยใช้ตำแหน่งของข้อมูลสูญหายเดิมคือ y_{22} และ y_{33} ได้ดังแสดงในตารางที่ 12 – 13

1) ประมาณค่าข้อมูลสูญหายด้วยวิธี Trajectory Mean

ตารางที่ 12 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ Trajectory Mean

เวลา (สัปดาห์)	หน่วยทดลอง				$\overline{y_{.j}^{IM}}$
	1	2	3	4	
ก่อนทดลอง	216	130	206	106	164.5
8 สัปดาห์	290	**191.25**	193	130	201.0625
16 สัปดาห์	340	206	**212**	144	225.5
24 สัปดาห์	275	220	223	150	217
32 สัปดาห์	294	209	226	173	225.5
$\overline{y_{i.}}$	283	191.25	212	140.6	

2) คำนวณค่าความผันแปรเฉลี่ย $\left(AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}\right)$

$\left(AV_2 = \overline{y_{.2}} - \overline{y_{.2}^{IM}}\right)$ = 204.33 – 201.0625 = 3.2675

$\left(AV_3 = \overline{y_{.3}} - \overline{y_{.3}^{IM}}\right)$ = 230 – 225.5 = 4.5

3) คำนวณค่าข้อมูลสูญหาย $\left(\hat{y}_{ij} = y_{ij}^{IM} + AV_j\right)$

$\left(\hat{y}_{22} = y_{22}^{IM} + AV_2\right)$ = 191.25 + 3.2675 = 194.5175

$\left(\hat{y}_{33} = y_{33}^{IM} + AV_3\right)$ = 212 + 4.5 = 216.5

ดังนั้นจะได้ชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean Trajectory ดังนี้

ตารางที่ 13 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290	**194.5175**	193	130
16 สัปดาห์	340	206	**216.5**	144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

1.6.2 วิธีCopyMean LOCF

การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean LOCF มีขั้นตอนดังนี้

1) คำนวณค่าข้อมูลสูญหายจากวิธีการ LOCF โดยการแทนที่ค่าข้อมูลสูญหายด้วยค่าก่อนหน้าที่ไม่ได้เป็นค่าข้อมูลสูญหาย

2) คำนวณค่าความผันแปรเฉลี่ยจากการนำค่าเฉลี่ยของข้อมูลที่มีค่าข้อมูลสูญหายลบด้วยค่าเฉลี่ยของชุดข้อมูลที่แทนที่ค่าข้อมูลสูญหายด้วยการประมาณค่าด้วยวิธีการ LOCF แล้ว $\left(AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}\right)$

3) คำนวณค่าข้อมูลสูญหายด้วยการนำข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ วิธีการ LOCF ไปบวกกับค่าความผันแปรเฉลี่ย $\left(\hat{y}_{ij} = y_{ij}^{IM} + AV_j\right)$

<u>ตัวอย่างที่ 4</u> การประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean LOCF โดยใช้ตัวอย่างจากชุดข้อมูลน้ำหนักผมที่สุ่มค่าข้อมูลสูญหายแล้วจากตัวอย่างที่ 2 ในตารางที่ 6 โดยใช้ตำแหน่งของข้อมูลสูญหายเดิมคือ y_{22} และ y_{33} ได้ดังแสดงในตารางที่ 14 - 15

1) ประมาณค่าข้อมูลสูญหายด้วยวิธี LOCF ได้ดังแสดงในตารางที่ 14

ตารางที่ 14 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ LOCF

เวลา (สัปดาห์)	หน่วยทดลอง				$\overline{y_{.j}^{IM}}$
	1	2	3	4	
ก่อนทดลอง	216	130	206	106	164.5
8 สัปดาห์	290	**130**	193	130	189.466
16 สัปดาห์	340	206	**193**	144	222.6
24 สัปดาห์	275	220	223	150	217
32 สัปดาห์	294	209	226	173	225.5

2) คำนวณค่าความผันแปรเฉลี่ย $\left(AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}\right)$

$\left(AV_2 = \overline{y_{.2}} - \overline{y_{.2}^{IM}}\right)$ = 204.33 – 189.466 = 14.864

$\left(AV_3 = \overline{y_{.3}} - \overline{y_{.3}^{IM}}\right)$ = 230 – 222.6 = 7.4

3) คำนวณค่าข้อมูลสูญหาย $\left(\hat{y}_{ij} = y_{ij}^{IM} + AV_j\right)$

$\left(\hat{y}_{22} = y_{22}^{IM} + AV_2\right)$ = 130 + 14.864 = 144.864

$\left(\hat{y}_{33} = y_{33}^{IM} + AV_3\right)$ = 193 + 7.4 = 200.4

ดังนั้นจะได้ชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธี CopyMean LOCF ดังนี้

ตารางที่ 15 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290	**144.864**	193	130
16 สัปดาห์	340	206	**200.4**	144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

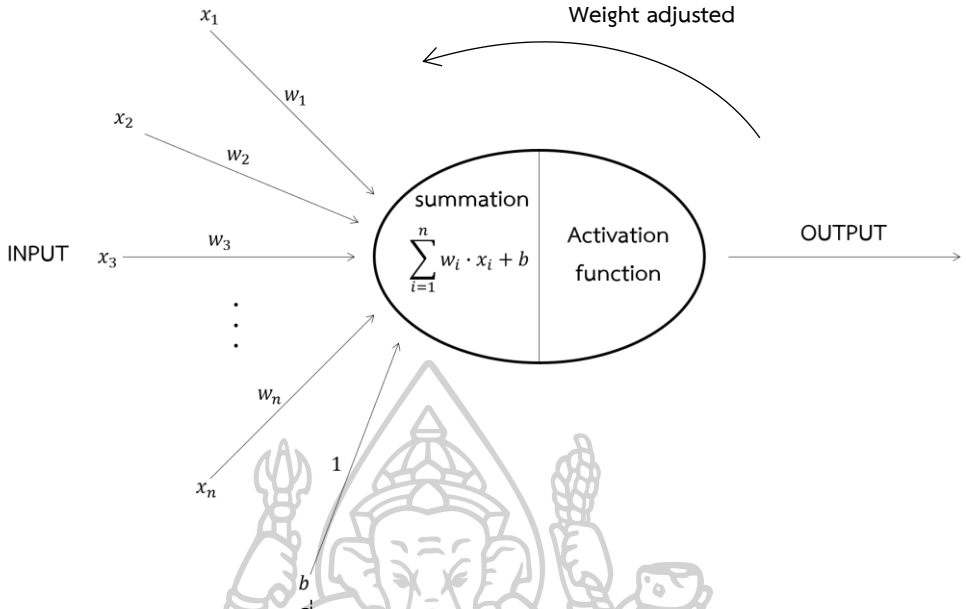
1.7 การประมาณค่าข้อมูลสูญหายด้วยวิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN)

Artificial Neural Networks (ANN) มีแนวคิดมาจากการทำงานของระบบประสาทของสิ่งมีชีวิต ซึ่งมีกระบวนการที่คล้ายกันคือการเรียนรู้จากเหตุการณ์ที่เกิดขึ้นก่อนหน้าเพื่อวิเคราะห์เหตุการณ์ที่เกิดขึ้นใหม่ โดยโครงข่ายประสาทเทียมนั้นสามารถนำมาประยุกต์ได้หลากหลายเช่น การวินิจฉัยโรค การตรวจลายนิ้วมือ การตรวจจับเสียงพูด ตรวจจับความผิดพลาดในกระบวนการทางเคมี นอกจากนี้ยังมีประโยชน์ในการนำมาใช้พยากรณ์และประมาณค่าข้อมูลสูญหายอีกด้วย (Gupta & Lam, 1996)

1.7.1 กระบวนการเรียนรู้ของโครงข่ายประสาทเทียม (Artificial Neural Network processes) มีขั้นตอนการทำงานของโครงข่ายประสาทเทียมมีอยู่ 7 ขั้นตอนดังนี้

1) นำเข้าเวกเตอร์ของข้อมูลตัวแปรอิสระ
2) คูณเวกเตอร์ของข้อมูลตัวแปรอิสระด้วยเซตของน้ำหนัก
3) ดำเนินการคำนวณค่าผลลัพธ์ด้วย activation function
4) ส่งกลับข้อมูลผลลัพธ์
5) คำนวณค่าความคลาดเคลื่อน (Error) เพื่อใช้ในการปรับค่าน้ำหนัก
6) ใช้ค่าน้ำหนักที่ปรับแล้วในกระบวนการรอบถัดไป
7) ทำการวนซ้ำในขั้นตอนที่ 1 ถึง 6(Shamdasani, 2017)

โดยขั้นตอนการทำงานของ Artificial Neural Network สามารถเขียนเป็นแผนภาพได้ดังนี้



ภาพที่ 8 กระบวนการ backpropagation

ในการสร้างโครงข่ายประสาทเทียมนั้นการเลือกค่าน้ำหนักนับเป็นส่วนสำคัญในการประมาณค่าผลลัพธ์ซึ่งการปรับค่าน้ำหนักไปเรื่อย ๆ ในแต่ละรอบของการทำซ้ำจะทำให้ได้ผลลัพธ์ของการประมาณค่าใกล้เคียงกับค่าจริงมากขึ้น แต่นอกจากค่าน้ำหนักแล้วการเลือกActivation function ที่เหมาะสมก็เป็นส่วนสำคัญเช่นกันในการประมาณค่า

1.7.2 Activation function ที่เลือกใช้ในการประมาณค่าข้อมูลสูญหายในข้อมูลแบบวัดซ้ำ

การเลือกใช้ Activation function ควรเลือกโดยคำนึงถึงค่าของตัวแปรควรจะสัมพันธ์กับ Activation function ในที่นี้เนื่องจากทราบว่าข้อมูลแบบวัดซ้ำนั้นตัวแปรตามมีลักษณะเป็นตัวแปรแบบต่อเนื่องที่ไม่ได้มีลักษณะซับซ้อนมากดังนั้นจึงเลือกใช้ฟังก์ชั่นเชิงเส้นเป็น Activation function

1.7.3 การประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมด้วยโปรแกรม R

การประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมด้วยโปรแกรม R นั้นจะทำโดยใช้แพ็คเกจที่มีชื่อว่า “Neuralnet” ซึ่งเป็นแพ็คเกจที่ใช้เพื่อสร้างโครงข่ายประสาทเทียม โดยจะมีฟังก์ชั่นสำคัญที่ใช้งานดังนี้

Neuralnet(formula, data, hidden, threshold, stepmax, rep, startweights, learningrate, lifesign, algorithm, err.fct, act.fct, linear.output, exclude)

โดยคำสั่งในฟังก์ชั่น ‘Neuralnet’ มีความหมายดังนี้

คำสั่ง	หมายถึง
formula	ตัวแบบที่ต้องการพยากรณ์
data	data frame ที่ประกอบด้วยตัวแปรที่จะนำมาสร้างโครงข่ายประสาทเทียม
hidden	เวกเตอร์ของจำนวนนิวรอนของ hidden layer แต่ละเลเยอร์
threshold	ค่าของการปรับฟังก์ชั่นความคลาดเคลื่อยและใช้เป็นจุดสิ้นสุกของกระบวนการวนซ้ำ
stepmax	จำนวนขั้นตอนที่มากที่สุดของการสร้างโครงข่ายประสาทเทียม
rep	จำนวนซ้ำของการทำซ้ำของการคำนวณโครงข่าย
startweights	เวกเตอร์ของค่าน้ำหนักเริ่มต้น หากใส่ค่า NULL ค่าน้ำหนักเริ่มต้นจะเป็นค่าสุ่ม
learningrate	ค่าของ learning rate ที่ใช้ในกระบวนการ backpropagation.
lifesign	คำสั่งพิมพ์กระบวนการในการสร้างค่าผลลัพธ์โดย 'none' คือไม่พิมพ์ 'minimal' คือพิมพ์กระบวนการอย่างย่อ และ 'full' คือพิมพ์กระบวนการทุกขั้นตอน
algorithm	กระบวนการที่ใช้ในการบค่าน้ำหนักในโครงข่ายประสาทเทียมประกอบด้วยฟังก์ชัน 'backprop' กระบวนการ backpropagation 'rprop+' และ 'rprop-' คือกระบวนการ resilient backpropagation ที่ดำเนินการและไม่ดำเนินการด้วย weight backtracking, 'sag' และ 'slr' คือกระบวนการปรับค่าความคลาดเคลื่อนเป็นค่าต่ำสุดสัมบูรณ์
err.fct	ฟังก์ชันอนุพันธ์ที่ใช้ในการคำนวณค่าความผิดพลาดโดย 'sse' และ 'ce' คือการคำนวณค่าความผิดพลาดด้วย sum of squared errors และ the cross-entropy ตามลำดับ
act.fct	Activation function หรือฟังก์ชันอนุพันธ์ที่ใช้ในการคำนวณค่าผลลัพธ์
linear.output	‘TRUE’ คือการแสดงค่าผลลัพธ์เป็นเชิงเส้นนอกจากนี้ใช้ ‘FALSE’
exclude	เวกเตอร์หรือเมทริกซ์ของน้ำหนักที่กำหนดไว้แล้วโดยค่าน้ำหนักนี้จะไม่ถูกปรับโดยกระบวนการของโครงจ่ายประสาทเทียม

ที่มา : (Fritsch, Guenther, & Wright, 2019)

<u>ตัวอย่างที่ 5</u> กระบวนการประมาณค่าข้อมูลสูญหายโดยใช้กระบวนการโครงข่ายประสาทเทียมโดยโปรแกรม R มีดังนี้

ใช้ตัวอย่างจากชุดข้อมูลน้ำหนักผมที่สุ่มค่าข้อมูลสูญหายแล้วจากตัวอย่างที่ 2 ในตารางที่ 6

1) นำเข้าชุดข้อมูลและตัดส่วนที่เป็นค่าข้อมูลสูญหายออก

```
z1 <- c(216,  290,  340,  275,  294)
z2 <- c(130,  NA,   206,  220,  209)
z3 <- c(206,  193,  NA,   223,  226)
z4 <- c(106,  130,  144,  150,  173)
zz <- data.frame(z1, z2, z3, z4)
zz2 <- zz[-2,-3]
zz3 <- zz[-3,-2]
```

2) ประมาณค่าข้อมูลสูญาหายโดยใช้คำสั่งจากฟังก์ชันใน package “Neuralnet” ดังนี้

2.1) สร้างฟังก์ชัน switch เพื่อใช้เป็น activation function

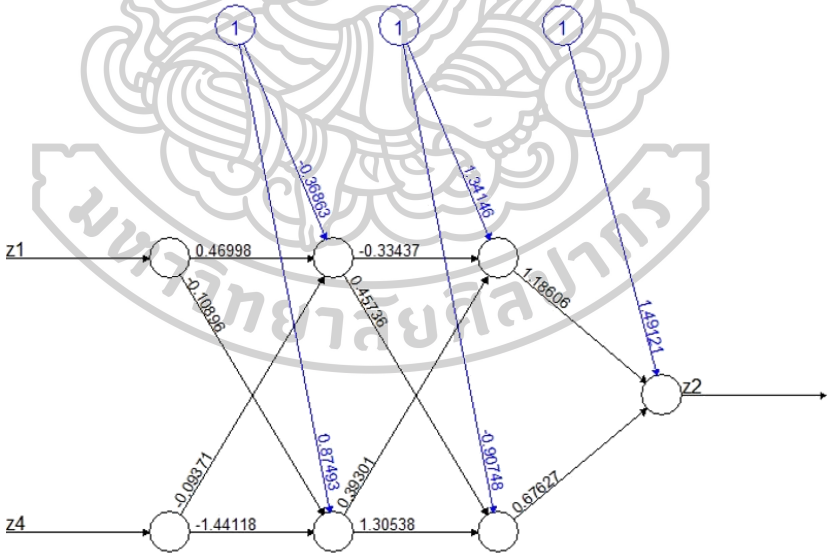
```
swt <- function(x) (2*x)/(1+exp(-x))
```

2.2) ประมาณค่าข้อมูลสูญหายโดยใช้ฟังก์ชัน “Neuralnet” โดยกำหนดให้คอลัมน์ที่มีค่าข้อมูลสูญหายเป็นตัวแปรตาม และคอลัมน์ที่ไม่มีค่าข้อมูลสูญหายเป็นตัวแปรอธิบาย และกำหนดให้โครงข่ายประกอบด้วย hidden layers 2 เลเยอร์ เลเยอร์ละ 2 นิวรอน ใช้กระบวนการในการปรับค่าน้ำหนักคือ กระบวนการ Backpropagration และ switch function เป็น Activation function กำหนดค่า learning rate เป็น 0.0001 แสดงค่าความผิดพลาดด้วยค่าความคลาดเคลื่อน กำลังสอง

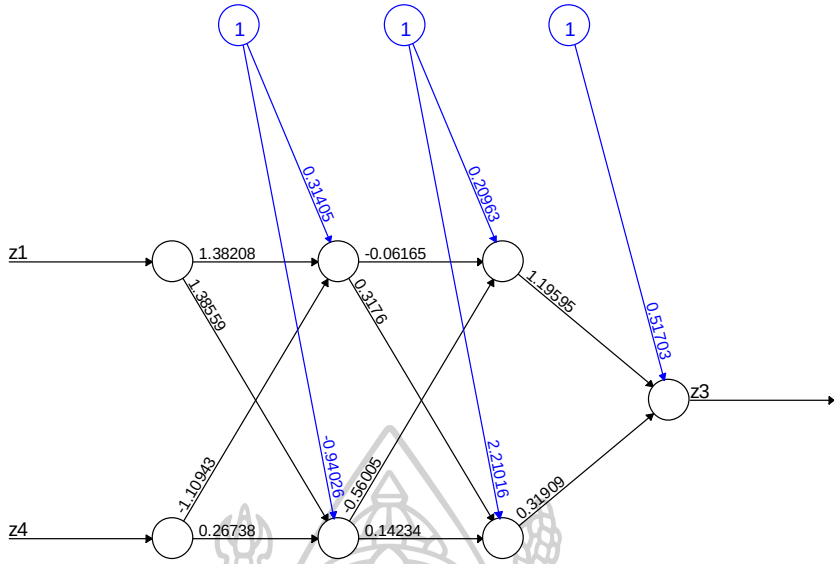
```
zn2 <- Neuralnet(z2~z1+z4,data = zz2, hidden = c(2,2),lifesign = 'minimal', rep = 1,
      algorithm='backprop', act.fct = swt, learningrate = 0.0001, err.fct = 'sse',
      linear.output=TRUE)

pred2 <- data.frame(z1=290,z4=130)

Predict2 <- compute(zn2,pred2)
```

```
Predict2$net.result

plot(zn2)

zn3 <- Neuralnet(z3~z1+z4, data = zz3,hidden = c(2,2),lifesign = 'minimal', rep
= 1,          algorithm='backprop',act.fct = swt, linear.output=TRUE,learningrate =
                    0.0001,err.fct = 'sse')

pred3 <- data.frame(z1=340, z4=144)

Predict3 <- compute(zn3, pred3)

Predict3$net.result

plot(zn3)
```

ผลลัพธ์ของค่าน้ำหนักและความคลาดเคลื่อนของโครงข่ายประสาทเทียมสำหรับการประมาณค่าข้อมูล สูญหายของ y_{22} และ y_{33} สามารถแสดงโดยแผนภาพได้ดังนี้



ภาพที่ 9 แผนภาพโครงข่ายประสาทเทียมสำหรับการประมาณค่าข้อมูลสูญหายของ y_{22}



ภาพที่ 10แผนภาพโครงข่ายประสาทเทียมสำหรับการประมาณค่าข้อมูลสูญหายของ y_{33}

ดังนั้นจะได้ชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมดังนี้

ตารางที่ 16 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียม

เวลา (สัปดาห์)	หน่วยทดลอง			
	1	2	3	4
ก่อนทดลอง	216	130	206	106
8 สัปดาห์	290	**153.36**	193	130
16 สัปดาห์	340	206	**220.18**	144
24 สัปดาห์	275	220	223	150
32 สัปดาห์	294	209	226	173

จากข้อมูลข้างต้นสามารถทำตารางสรุปค่าประมาณข้อมูลสูญหายได้ดังในตารางที่ 17

ตารางที่ 17 ข้อมูลจากตัวอย่างที่ 2 ที่ประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีCopyMean และวิธีโครงข่ายประสาทเทียม

ข้อมูลสูญหาย		วิธีการประมาณค่าข้อมูลสูญหาย			
ตำแหน่ง	ค่าจริง	MS	CT	CL	ANN
$\hat{y}_{22}$	146	186.33	194.52	144.86	153.36
$\hat{y}_{33}$	218	239.17	216.50	200.40	220.18

1.8 เกณฑ์ที่ใช้ในการประเมิน

การประมาณค่าข้อมูลสูญหายเสร็จสิ้นแล้วการพิจารณาว่าวิธีการใดเป็นวิธีการที่ใช้ในการประมาณค่าข้อมูลสูญหายได้ดีที่สุด สามารถทดสอบโดยใช้เกณฑ์ในการประเมิน โดยในที่นี้ใช้เกณฑ์ในการประเมิน 3 วิธีดังนี้

1.8.1 ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute deviation : MAD)

ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (MAD) เป็นวิธีการที่ใช้วัดความแตกต่างระหว่างค่าจริงและค่าพยากรณ์หรือความแตกต่างระหว่างค่าสังเกตกับค่ากลางของข้อมูลเช่น ค่าเฉลี่ย ค่ามัธยฐาน หรือ ค่าฐานนิยม โดยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยจะแสดงค่าในรูปของระยะห่างระหว่างค่าพยากรณ์กับค่าจริง โดยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยจะมีค่าไม่เป็นลบเสมอและค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยที่เท่ากับศูนย์จะหมายความว่าค่าพยากรณ์มีค่าตรงกับค่าจริงทุกค่า เนื่องจากค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยแสดงถึงความผิดพลาดของการประมาณค่า ดังนั้นโดยทั่วไปแล้ววิธีการพยากรณ์ที่ให้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยมีค่าต่ำ จะมีความเหมาะสมกว่าวิธีการพยากรณ์ที่ให้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยที่สูงกว่า แต่วิธีการนี้มีข้อเสียคือไม่สามารถวัดทิศทางการกระจายของข้อมูลได้ (LUND & LUND, 2018a)

ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยสามารถแสดงได้ดังสมการ

เมื่อกำหนดให้ N คือ จำนวนชุดข้อมูลทั้งหมด

M คือ จำนวนค่าข้อมูลสูญหาย

h คือ index ของชุดข้อมูลที่ h

k คือ index ของค่าข้อมูลสูญหายที่ k

y_{ijk} คือ ค่าจริงของค่าสังเกต ij ที่เก็บจากตัวแปรที่ i เวลาที่ j ที่เป็นค่าข้อมูลสูญหายค่าที่ k

$\hat{y}_{ijk}$ คือ ค่าพยากรณ์ของค่าสังเกต ij ที่เป็นค่าสูญหายค่าที่ k

$$MAD = \frac{\sum_{h=1}^{N}\sum_{k=1}^{M}\left|\hat{y}_{ijk} - y_{ijk}\right|}{N \times M}$$

ตัวอย่างที่ 6 การคำนวณค่าค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย

ใช้ตัวอย่างจากชุดข้อมูลน้ำหนักผม และข้อมูลที่ประที่ประมาณค่าข้อมูลสูญหายแล้ว

ตัวอย่างการคำนวณค่า MAD จากการประมาณค่าข้อมูลสูญหาย

ตารางที่ 18 ค่า MAD สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม

วิธีการประมาณค่าข้อมูลสูญหาย	MAD
MS	$\frac{\left\|186.33-146\right\|+\left\|239.17-218\right\|}{1\times 2}=30.75$
CT	$\frac{\left\|194.52-146\right\|+\left\|216.50-218\right\|}{1\times 2}=25.01$
CL	$\frac{\left\|144.86-146\right\|+\left\|200.40-218\right\|}{1\times 2}=9.37$
ANN	$\frac{\left\|153.36-146\right\|+\left\|220.18-218\right\|}{1\times 2}=4.77$

เมื่อใช้ค่า MAD เป็นเกณฑ์ที่ใช้ในการประเมินค่าข้อมูลสูญหาย ผลลัพธ์ที่ได้คือการประมาณค่าข้อมูลสูญหายด้วยวิธีโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เหมาะสมที่สุด

1.8.2 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย *(*Root Mean Square deviation *: RMSD)*

รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square deviation : RMSD) เป็นวิธีการที่ใช้วัดความแตกต่างระหว่างค่าจริงและค่าพยากรณ์หรือความแตกต่างระหว่างค่าสังเกตกับค่ากลางของข้อมูลโดยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยจะแสดงค่าในรูปของรากที่สองของค่าเฉลี่ยของผลต่างกำลังสองระหว่างค่าจริงและค่าพยากรณ์หรือที่เรียกว่าส่วนเหลือซึ่งวิธีการประเมินด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสามารถใช้ได้กับทั้งประเมินความแตกต่างระหว่างค่าจริงกับค่าพยากรณ์ หรือใช้ประเมินความแตกต่างระหว่างข้อมูล 2 ชุดก็ได้ (Robinson, 2017)

วิธีการคำนวณค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยมีดังนี้

$$RMSD = \sqrt{\frac{\sum_{h=1}^{N}\sum_{k=1}^{M}\left(\hat{y}_{ijk} - y_{ijk}\right)^2}{N \times M}}$$

รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าไม่เป็นลบเสมอและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่เท่ากับศูนย์จะหมายความว่าค่าพยากรณ์มีค่าตรงกับค่าจริงทุกค่า เนื่องจากค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยแสดงถึงความผิดพลาดของการประมาณค่า ดังนั้นโดยทั่วไปแล้ววิธีการพยากรณ์ที่ให้ค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าต่ำ จะมีความเหมาะสมกว่าวิธีการพยากรณ์ที่ให้ค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่สูงกว่าเช่นเดียวกับ MAD แต่อย่างไรก็ตามวิธีการประเมินด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย จะไม่เหมาะสมหากต้องการใช้เปรียบเทียบข้อมูล 2 ชุดที่มีลักษณะของข้อมูลต่างกันมาก เนื่องจากวิธีการประเมินด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยการประเมินขึ้นกับ scale ของข้อมูลด้วย นอกจากนี้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยยังมีความไวต่อค่านอกกลุ่ม กล่าวคือค่านอกกลุ่มจะทำให้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าสูงกว่าปรกติเป็นอย่างมาก วิธีการประเมินนี้จึงไม่เหมาะสมกับข้อมูลที่มีค่านอกกลุ่ม (Stephanie, 2016)

ตัวอย่างที่ 7 การคำนวณค่า RMSD จากการประมาณค่าข้อมูลสูญหาย

ตารางที่ 19 ค่า RMSD สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม

วิธีการประมาณค่าข้อมูลสูญหาย	RMSD
MS	$\sqrt{\frac{(186.33-146)^2+(239.17-218)^2}{1\times 2}}=32.21$
CT	$\sqrt{\frac{(194.52-146)^2+(216.50-218)^2}{1\times 2}}=34.33$
CL	$\sqrt{\frac{(144.86-146)^2+(200.40-218)^2}{1\times 2}}=12.47$
ANN	$\sqrt{\frac{(153.36-146)^2+(220.18-218)^2}{1\times 2}}=7.03$

เมื่อใช้ค่า RMSD เป็นเกณฑ์ที่ใช้ในการประเมินค่าข้อมูลสูญหาย ผลลัพธ์ที่ได้คือการประมาณค่าข้อมูลสูญหายด้วยวิธีโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เหมาะสมที่สุด

1.8.3 ค่าความเอนเอียง (Bias)

ค่าความเอนเอียง (Bias) เป็นวิธีการวัดค่าความคลาดเคลื่อนในการพยากรณ์โดยพิจารณาทิศทางของข้อมูลร่วมด้วย การคำนวณค่าความเอนเอียงของหน่วยสังเกตแต่ละค่าสามารถทำได้โดยการหาค่าความแตกต่างระหว่างค่าพยากรณ์และค่าจริงซึ่งผลต่างนี้อาจมีเครื่องหมายเป็นลบหรือบวกก็ได้ และในส่วนของการคำนวณค่าความเอนเอียงสำหรับการประมาณค่าหลาย ๆ ค่าสามารถทำได้โดยการหาค่าความเอนเอียงเฉลี่ยแต่ละค่า หากค่าความเอนเอียงนี้มีค่าเท่ากับศูนย์จะเรียกว่าไม่มีความเอนเอียง (Unbiased) หากการประมาณค่ามีความค่าความเอนเอียงหมายความว่ากระบวนการประมาณค่าอาจมีความผิดพลาดซึ่งจะทำให้การใช้วิธีการประมาณค่านั้น ๆ ประมาณค่าข้อมูลได้คลาดเคลื่อน (Berman, 2019)

วิธีการคำนวณค่าความเอนเอียงมีดังนี้

$$MAD = \frac{\sum_{h=1}^{N}\sum_{k=1}^{M}\left(\hat{y}_{ijk} - y_{ijk}\right)}{N \times M}$$

อย่างไรก็ตามค่าความเอนเอียงจะเป็นวิธีการประเมินที่ไม่ดีหากค่าความเอนเอียงของข้อมูลแต่ละค่ามีค่ามากแต่มีทิศทางตรงกันข้ามกัน โดยจะทำให้ไม่เห็นค่าความคลาดเคลื่อนของการประมาณค่าข้อมูลทั้งชุด (Markgraf, 2018) ซึ่งอาจพิจารณาวิธีการประเมินด้วย MAD หรือ RMSD ร่วมด้วย

ตัวอย่างที่ 8 การคำนวณค่า Bias จากการประมาณค่าข้อมูลสูญหาย

ตารางที่ 20 ค่า Bias สำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม

วิธีการประมาณค่าข้อมูลสูญหาย	Bias
MS	$\frac{(186.33-146)+(239.17-218)}{1\times 2} = 30.75$
CT	$\frac{(194.52-146)+(216.50-218)}{1\times 2} = 23.51$
CL	$\frac{(144.86-146)+(200.40-218)}{1\times 2} = -9.37$
ANN	$\frac{(153.36-146)+(220.18-218)}{1\times 2} = 4.77$

เมื่อใช้ค่า RMSD เป็นเกณฑ์ที่ใช้ในการประเมินค่าข้อมูลสูญหาย ผลลัพธ์ที่ได้คือการประมาณค่าข้อมูลสูญหายด้วยวิธีโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เหมาะสมที่สุด เช่นเดียวกับเมื่อใช้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยในการประเมิน

จากการคำนวณค่าที่ใช้ในการประเมินวิธีการประมาณค่าข้อมูลสูญหายทั้ง 3 วิธีสามารถสรุปผลการประมาณค่าข้อมูลสูญหายได้ดังตารางที่ 21

ตารางที่ 21 ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายสำหรับการประมาณค่าข้อมูลสูญหายด้วยวิธีวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีCopyMean และวิธีโครงข่ายประสาทเทียม

เกณฑ์การประเมิน	วิธีการประมาณค่าข้อมูลสูญหาย			
	MS	CT	CL	ANN
MAD	30.75	25.01	9.37	4.77
RMSD	32.21	34.33	12.47	7.03
Bias	30.75	23.51	-9.37	4.77

จากตารางข้างต้นจะเห็นว่าวิธีการโครข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพมากที่สุดรองลงมาคือวิธีการ CopyMean LOCF คือวิธีการ CopyMean LOCF เป็นวิธีการที่ใช้ประมาณค่าข้อมูลสูญหายแล้วมีค่าการประเมินโดยใช้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยในการประเมิน และค่าความเอนเอียงต่ำที่สุด

2. งานวิจัยที่เกี่ยวข้อง

Gupta และ Lam (1996) ศึกษาการประมาณค่าข้อมูลสูญหายและการจัดกลุ่มโดยใช้วิธีโครงข่ายประสาทเทียม ซึ่งการจัดกลุ่มเป็นเครื่องมือที่สำคัญอย่างนี้ที่ช่วยในการตัดสินใจและสามารถนำมาประยุกต์ใช้ในการแก้ปัญหาทางธุรกิจ หรือพยากรณ์จำนวนครั้งการใช้บัตรเครดิตเป็นต้น ในหลาย ๆ งานวิจัยประยุกต์ใช้วิธีโครงข่ายประสาทเทียมในการจัดกลุ่มของตัวแปร ซึ่งพบว่าวิธีการโครงข่ายประสาทเทียมมีความเหมาะสมกว่าวิธีการทางสถิติ และในงานวิจัยนี้ Gupta และ Lam สนใจที่จะนำวิธีการโครงข่ายประสาทเทียมมาใช้ในการประมาณค่าข้อมูลสูญหาย ค่าข้อมูลสูญหายเป็นปัญหาที่พบเจอได้บ่อยครั้งในการวิเคราะห์ข้อมูลทางสถิติ ซึ่งข้อมูลสูญหายจะก่อให้เกิดปัญหาตามมาได้แก่ 1) ขนาดตัวอย่างลดลงซึ่งเป็นผลให้นัยสำคัญทางสถิติสำหรับการสรุปผลลดลง 2) ข้อมูลสูญหายจะทำให้ความสามารถในการพยากรณ์และประมาณค่าลดลง 3) อาจมีสารสนเทศที่สำคัญอยู่ในส่วนของข้อมูลที่เป็นค่าข้อมูลสูญหาย ดังนั้นการประมาณค่าข้อมูลสูญหายจึงเป็นทางเลือกหนึ่งที่ควรทำซึ่งผู้วิจัยได้สนใจใช้วิธีโครงข่ายประสาทเทียมมาประมาณค่าข้อมูลสูญหาย วิธีโครงขายประสาทเทียมมีแนวคิดมาจากระบบประสาทของมนุษย์ โดยวิธีโครงข่ายประสาทเทียมมีกระบวนการทำงานโดยการเรียนรู้จากสภาพแวดล้อมหรือประสบการณ์ก่อนหน้าแล้วนำมาพยากรณ์ผลลัพธ์ที่สนใจและสามารถนำไปประยุกต์ใช้ในงานได้หลากหลายเช่นใช้ในการตรวจจับลายมือ ลายนิ้วมือ หรือเสียงพูด ตรวจสอบความผิดพลาดในกระบวนการทางเคมี และยังสามารถใช้กระบวนการ

backpropagation ในโครงข่ายประสาทเทียมมาใช้ประโยชน์ในการแบ่งกลุ่มของตัวแปรได้อีกด้วย ผู้วิจัยได้ใช้วิธีการโครงข่ายประสาทเทียมมาใช้ในการประมาณค่าข้อมูลสูญหายในข้อมูลที่มีลักษณะเป็นเชิงกลุ่มและนำผลลัพธ์มาเปรียบเทียบกับวิธีการอื่น ๆ ผลลัพธ์ปรากฏว่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายที่มีลักษณะเป็นเชิงกลุ่ม

Bingham, Stemmler, Peterson, และ Graber (1998) ได้ศึกษาการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำ โดยศึกษาการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำโดยแบ่งการทดลองออกเป็นสองส่วน ส่วนแรกคือการจำลองแบบครั้งเดียว คือจะสุ่มข้อมูลโดยใช้ค่าพารามิเตอร์จากชุดข้อมูลจริงจำนวน 183 ชุดข้อมูล แต่ละชุดข้อมูลสุ่มตัวอย่างขนาด 5 หน่วยตัวอย่างและวัดซ้ำ 4 ครั้ง โดยมีลักษณะของข้อมูล คือเป็นกราฟเส้นตรง กราฟกำลังพาราโบลิก กราฟรูปตัวเอส และกราฟไซน์ ตามลำดับ ส่วนที่สองจะใช้วิธีการจำลองข้อมูลแบบมอนติคาร์โลโดยกำหนดจำนวนการวนซ้ำ 1000 รอบ โดยกำหนดรูปร่างของข้อมูลคือกราฟเส้นตรง กราฟกำลังตามแนวตั้ง กราฟกำลังตามแนวนอน พาราโบลิก กราฟรูปตัวเอส และกราฟไซน์ แต่ละฟังก์ชั่นจะสุ่มขนาดตัวอย่างจำนวน 120, 240 และ 480 ค่าโดยแต่ละชุดข้อมูลสุ่มตัวอย่างขนาด 5 หน่วยตัวอย่างและวัดซ้ำ 4 ครั้งในแต่ละหน่วยตัวอย่าง ในการทดลองทั้งสองส่วนจะสุ่มค่าข้อมูลสูญหายตามสัดส่วนคือ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 และ0.7 ตามลำดับจากนั้นคำนวณค่าความคลาดเคลื่อนมาตรฐานและค่าโมเมนต์ ผลลัพธ์ปรากฏว่าการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าประมาณใกล้เคียงกับค่าจริงแม้กระทั้งเมื่อขนาดตัวอย่างเกิน 100 ค่าและรูปร่างของข้อมูลไม่ได้มีอิทธิพลต่อการประมาณค่าข้อมูลสูญหาย

Rubin, Witkiewitz, Andre, และ Reilly (2007) ศึกษาวิธีการดำเนินการเกี่ยวกับข้อมูลสูญหายในงานวิจัยเกี่ยวกับระบบประสาทโดยใช้กรณีศึกษาคือข้อมูลชุด “Don't Throw the Baby Rat out with the Bath Water” ค่าข้อมูลสูญหายเป็นปัญหาที่มักพบในการศึกษาด้านระบบประสาท ซึ่งในงานวิจัยส่วนใหญ่มักจะมีวิธีการแก้ปัญหาเรื่องค่าข้อมูลสูญหายด้วยการตัดค่าข้อมูลสูญหายทิ้ง ซึ่งวิธีการตัดค่าข้อมูลสูญหายส่งผลให้ผลลัพธ์ในการวิเคราะห์ข้อมูลเกิดความเอนเอียง ทำให้กำลังการทดสอบทางสถิติลดลง ทำให้ผลลัพธ์ในการประมาณค่า และการสรุปผลผิดพลาด ดังนั้นจึงไม่ควรลบค่าสังเกตที่มีค่าข้อมูลสูญหายออก และดำเนินการประมาณค่าข้อมูลสูญหายต่อไป วิธีการดำเนินการกับค่าข้อมูลสูญหายมีด้วยกันหลายวิธี โดย Rubin และคณะสนใจศึกษาวิธีการที่มีประสิทธิภาพกับค่าข้อมูลสูญหาย และคาดว่าการประมาณค่าข้อมมูลสูญหายเป็นวิธีการที่ดีกว่าการตัดค่าข้อมูลสูญหายทิ้ง เช่นวิธี listwise deletion ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลตามคาบเวลาที่เก็บผลกระทบของของ

ยาลดอาการอยากอาหารต่อการบริโภคอาหารของหนูจำนวน 17 ตัวโดยสุ่มค่าข้อมูลสูญหายจากตัวอย่างจำนวน 1, 2, 3, 4, 5 และ 10 เปอร์เซ็นต์ตามลำดับ จากนั้นดำเนินการกับค่าข้อมูลสูญหายด้วยวิธีการ 4 วิธีการ คือ listwise deletion, วิธีการแทนที่ด้วยค่าเฉลี่ย, การวิเคราะห์การถดถอย และวิธีค่าคาดหวังสูงสุด (EM) บนข้อมูลที่สุ่มค่าข้อมูลสูญหายทั้ง 6 ชุดแล้วคำนวณค่า P-values, ขนาดการทดสอบ และค่าปัจจัยของเบส์ ผลลัพธ์ของการทดลองคือวิธีการ listwise deletion เป็นวิธีที่มีประสิทธิภาพน้อยกว่าวิธีอื่นและวิธีการประมาณค่าข้อมูลสูญหายวิธี EM และวิธีการวิเคราะห์การถดถอยเป็นวิธีการที่ใช้ได้ดีกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย เมื่อสุ่มข้อมูลสูญหาย 10 เปอร์เซ็นต์ อย่างไรก็ดีผู้วิจัยแนะนำให้หลีกเลี่ยงการตัดค่าข้อมูลสูญหายทิ้งและประมาณค่าข้อมูลสูญหายแทนจะเป็นวิธีการที่เหมาะสมกว่า

Genolini, Lacombe, cochard, และ Subtil (2016) ได้ศึกษาการประมาณค่าข้อมูลสูญหายแบบทางเดียวในข้อมูลตามคาบเวลา ด้วยวิธีการ CopyMean ซึ่งเป็นวิธีการที่คิดขึ้นใหม่ โดยการศึกษาตามคาบเวลาเป็นการศึกษาโดยการวัดตัวแปรเดิมซ้ำ ๆ กันในต่างช่วงเวลาซึ่งการศึกษาตามคาบเวลานี้ก็สามารถพบค่าข้อมูลสูญหายได้เช่นเดียวกับการศึกษาประเภทอื่น โดยลักษณะของข้อมูลสูญหายสามารถแบ่งออกได้เป็น 3 ประเภทได้แก่ 1) การสูญหายแบบสุ่มสมบูรณ์ (MCAR) มีลักษณะคือความน่าจะเป็นที่จะเกิดค่าข้อมูลสูญหายเป็นอิสระกับตัวแปรอื่น ๆ 2) การสูญหายแบบสุ่ม (MAR) มีลักษณะคือความน่าจะเป็นที่จะเกิดค่าข้อมูลสูญหายขึ้นกับเฉพาะตัวแปรที่เก็บข้อมูลมาแล้ว 3) การสูญหายแบบไม่สุ่ม (MNAR) มีลักษณะที่มีความน่าจะเป็นซึ่งจะเกิดค่าข้อมูลสูญหายขึ้นกับทั้งตัวแปรที่เก็บข้อมูลมาแล้วและตัวแปนที่ยังไม่ได้เก็บข้อมูลมาก่อน เมื่อต้องการวิเคราะห์ข้อมูลเชิงสถิติสำหรับการศึกษาตามคาบเวลาพบว่าตัวแบบทางสถิติสำหรับการศึกษานี้ โดยการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดในตัวแบบผสม เป็นวิธีที่มีความแกร่งสำหรับลักษณะข้อมูลสูญหายแบบ MAR อย่างไรก็ตามเมื่อลักษณะข้อมูลสูญหายเป็นแบบ MNAR หรือเมื่อการวิเคราะห์ข้อมูลมีความไวต่อข้อสมมติเบื้องต้นทางสถิติตัวแบบ selection และตัวแบบ pattern mixture จะเหมาะสมสำหรับข้อมูลในลักษณะนี้ Genolini และคณะสนใจพิจารณาในสถานการณ์ที่ต้องการประมาณค่าเฉพาะในส่วนย่อยที่สนใจ เช่นการพยากรณ์ความดันเลือดในระหว่างการผ่าตัดที่อุปกรณ์วัดเสียหาย, ระยะเวลาในการรอรถไฟใต้ดิน หรือการพยากรณ์อากาศเป็นต้น และการตัดสินใจดังนั้น Genolini และคณะจึงใช้การประมาณค่าครั้งเดียวเนื่องจาก 1) ตัวแปรส่วนใหญ่มีลักษะเป็นแบบไม่ใช้พารามิเตอร์ดังนั้นทั้งแนวโน้มของข้อมูลและการแจกแจงของข้อมูลจะเปลี่ยนไปตามช่วงเวลาโดยเฉพาะอย่างยิ่งเมื่อข้อมูลตามคาบเวลาไม่เป็นพาราเมตริก ดังนั้น Genolini และคณะจึงใช้วิธีการทั่วไปแทนที่จะใช้วิธีที่ใช้

พื้นฐานจากตัวแบบหรือการวิเคราะห์ 2 ขั้น 2) ในบางกรณีการประมาณค่าข้อมูลสูญหายมีความสำคัญมากกว่าการประมาณค่าพารามิเตอร์ในตัวแบบ และ 3) สำหรับวิธีการทางสถิติบางวิธีการจะเหมาะสมสำหรับการวิเคราะห์บนข้อมูลที่ไม่มีค่าสูญหายเท่านั้น ในงานวิจัยนี้ Genolini และคณะได้เสนอวิธีการประมาณค่าข้อมูลสูญหาย 3 วิธีได้แก่ 1) วิธีการประมาณค่าตามขวางเป็นวิธีการประมาณค่าข้อมูลสูญหายโดยใช้สารสนเทศจากข้อมูลในเวลาที่ t เท่านั้นในการประมาณค่าข้อมูลสูญหาย 2) วิธีการประมาณค่าตามคาบเวลา เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ใช้ข้อมูลจากค่าสังเกตที่ i เท่านั้นในการประมาณค่าข้อมูลสูญหายและ 3) วิธีการผสมระหว่างการประมาณค่าตามขวางและการประมาณค่าตามคาบเวลา เป็นวิธีการที่ใช้สารสนเทศจากทั้งเวลาและหน่วยตัวอย่างในการประมาณค่าข้อมูลสูญหายร่วมกันโดยเสนอวิธี CopyMean ในการประมาณค่าข้อมูลสูญหายและพบว่าวิธีการนี้มีประสิทธิภาพมากกว่าวิธีการอื่น ในเกือบทุกสถานการณ์

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยเรื่องการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองได้แบ่งวิธีการดำเนินงานวิจัยออกเป็น 2 ส่วนได้การประยุกต์กับข้อมูลจริง และข้อมูลที่ได้โดยการจำลอง จากนั้นทำการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม และศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการประเมินด้วยค่า MAD RMSD และ Bias

1. ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่นำมาใช้งานวิจัยแบ่งเป็นข้อมูลจริงและข้อมูลที่ได้จากการจำลองดังนี้

1.1 ข้อมูลจากชุดข้อมูลจริงเป็นชุดข้อมูลที่เป็นข้อมูลแบบวัดซ้ำทั้งหมด 3 ชุดข้อมูลดังนี้

1.1.1 ข้อมูลชุด Drug Effect (Winer, 1962)

1.1.2 ข้อมูลชุด Skydive (Singley, Hale, & Russell, 2012)

1.1.3 ข้อมูลชุด Fecal Fat (Vittinghoff, Glidden, Shiboski, & McCulloch, 2012)

1.2 การจำลองข้อมูล

ข้อมูลจากการจำลองจะจำลองข้อมูลโดยกำหนดให้ชุดข้อมูลมีการแจกแจงปรกติพหุ 4 ตัวแปร (k=4) โดยมีขนาดตัวอย่างเท่ากับ 5 (n=5) และมีเวกเตอร์ค่าเฉลี่ยคือ [20,20,20,20]' มีค่าความแปรปรวนคือ 25 เท่ากัน และค่าสัมประสิทธ์สหสัมพันธ์ของข้อมูลแต่ละชุดคือ 0, 0.3, 0.5, 0.7, 0.9 ตามลำดับ

2. เครื่องมือที่ใช้ในการวิจัย

โปรแกรม R เวอร์ชั่น 3.6.1

3. สถิติที่ใช้ในการวิจัย

3.1 การประมาณค่าข้อมูลสูญหาย

ในการประมาณค่าข้อมูลสูญหายจะใช้การประมาณด้วยวิธีการทางสถิติดังต่อไปนี้

3.1.1 วิธีการแทนที่ด้วยค่าเฉลี่ย (Mean Substitution : MS)

3.1.2 วิธี CopyMean (CopyMean : CM) โดยแบ่งเป็น

3.1.2.1 วิธี CopyMean Trajectory (CopyMean Trajectory : CT)

3.1.2.2 วิธี CopyMean LOCF (CopyMean Trajectory :CL)

3.1.3 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network : ANN)

3.2 การประเมินประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย

ในการประเมินประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายจะพิจารณาจากสถิติดังต่อไปนี้

3.2.1 ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute deviation : MAD)

3.2.2 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square deviation : RMSD)

3.2.3 ค่าความเอนเอียง (Bias)

4. ขั้นตอนการจำลองข้อมูล

งานวิจัยนี้มีขั้นตอนการจำลองข้อมูลดังนี้

4.1 กำหนดขนาดตัวอย่างคือ n = 5

4.2 กำหนดให้ตัวแปรตามมีการแจกแจงปรกติพหุ 4 ตัวแปรโดยมีฟังก์ชั่นความน่าจะเป็นดังนี้

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

เมื่อ $\boldsymbol{\mu}$ คือ เวกเตอร์ค่าเฉลี่ย

$\boldsymbol{\Sigma}$ คือ เมทริกซ์ความแปรปรวนความแปรปรวนร่วม

โดยกำหนดพารามิเตอร์ดังต่อไปนี้

1) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปร โดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

2) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปร โดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.3;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

3) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปร โดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.5;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

4) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปร โดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.7;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

5) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปร โดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.9;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

4.3 จำลองข้อมูลซ้ำในแต่ละพารามิเตอร์จำนวน 1000 รอบ

5 วิธีการวิเคราะห์ข้อมูล

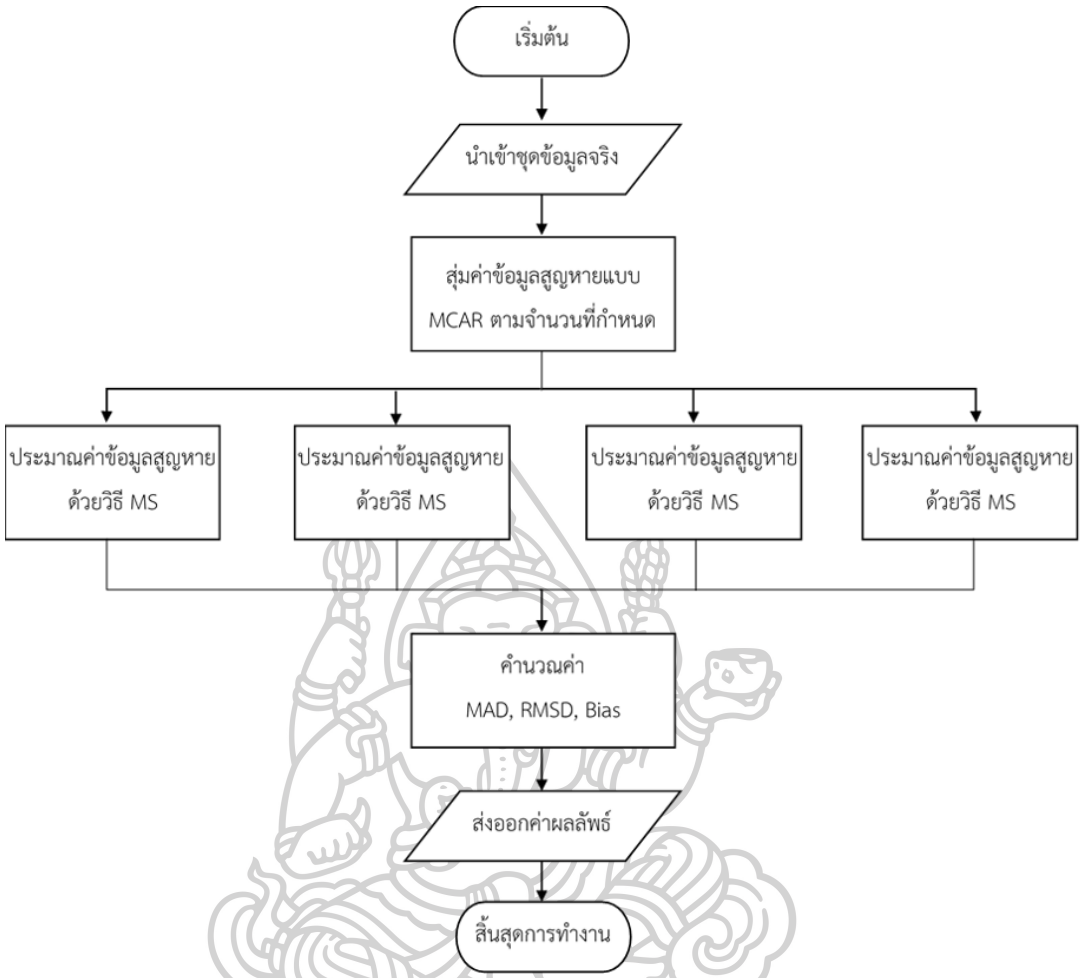
5.1 จำลองข้อมูลตามขั้นตอนการจำลองข้อมูลในขั้นตอนที่ 4.1 - 4.3

5.2 สุ่มค่าข้อมูลสูญหายทั้งจากในชุดข้อมูลจริงและชุดข้อมูลจำลอง จำนวน 1, 2 และ 3 ค่า ตามลำดับ โดยกำหนดลักษณะของข้อมูลสูญหายเป็นแบบการสูญหายอย่างสุ่มสมบูรณ์ (Missing Completely at Random : MCAR)

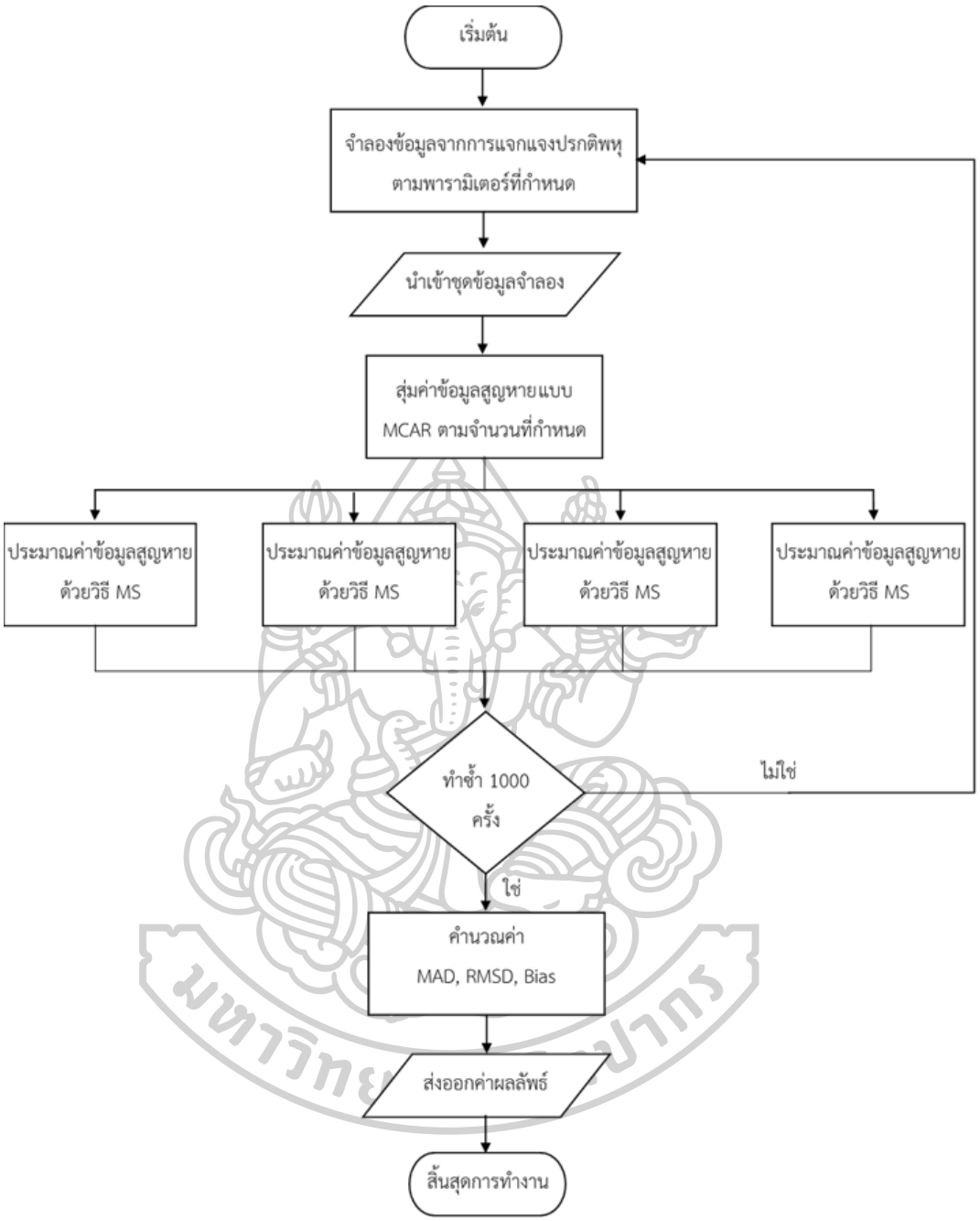
5.3 ประมาณค่าข้อมูลสูญหายด้วยวิธีการ แทนที่ด้วยค่าเฉลี่ย วิธี CopyMean และวิธีโครงข่ายประสาทเทียม ในข้อมูลทุกชุด

5.4 ประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยเกณฑ์การประเมินโดยเปรียบเทียบ ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และ ค่าความเอนเอียงของวิธีการประมาณค่าข้อมูลสูญหายทั้ง 3 วิธี

ขั้นตอนการดำเนินงานบนชุดข้อมูลจริงและชุดข้อมูลจำลองสามารถแสดงได้ในแผนภาพที่ 11 และ 12 ตามลำดับดังนี้



ภาพที่ 11 แผนภาพแสดงลำดับขั้นตอนการทำงานบนชุดข้อมูลจริง



ภาพที่ 12 แผนภาพแสดงลำดับขั้นตอนการทำงานบนชุดข้อมูลจำลอง

บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองเมื่อประมาณค่าข้อมูลสูญหายโดยใช้วิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean และวิธีโครงข่ายประสาทเทียม และเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายโดยใช้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าความเอนเอียงเป็นเกณฑ์ในการประเมินประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายในแต่ละสถานการณ์

ในการศึกษานี้ได้ทำการศึกษาโดยแบ่งผลการวิจัยออกเป็นสองส่วนคือส่วนที่ 1 คือผลการวิจัยจากชุดข้อมูลจริง และส่วนที่ 2 คือผลการวิจัยจากชุดข้อมูลจำลองโดยจำลองข้อมูลจากการแจกแจงแบบปรกติพหุในพารามิเตอร์ที่แตกต่างกัน 3 พารามิเตอร์

ส่วนที่ 1 **ผลการวิจัยโดยใช้ชุดข้อมูลจริง**

ในชุดข้อมูลจริงประกอบด้วยข้อมูลชุด Drug Effect ข้อมูลชุด Skydive และ ข้อมูลชุด Fecal Fat มีผลการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายดังนี้

1.1 ข้อมูลชุด Drug Effect

ข้อมูลชุดนี้ประกอบไปด้วย k = 4, n = 5 ข้อมูลที่ได้ถูกนำมาหาค่าของสถิติพรรณาจะได้เวกเตอร์ค่าเฉลี่ย เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วม และเมทริกซ์สหสัมพันธ์ของข้อมูลชุด Drug Effect แสดงดังนี้

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 26.4 & 25.6 & 15.6 & 32 \end{bmatrix}'$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 51.20 & 35.47 & 19.47 & 46.00 \\ 35.47 & 28.53 & 10.53 & 31.33 \\ 19.47 & 10.53 & 9.87 & 18.00 \\ 46.00 & 31.33 & 18.00 & 42.67 \end{bmatrix}$$

$$\hat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & 0.93 & 0.87 & 0.98 \\ 0.93 & 1 & 0.63 & 0.90 \\ 0.87 & 0.63 & 1 & 0.88 \\ 0.98 & 0.90 & 0.88 & 1 \end{bmatrix}$$

จากการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำในข้อมูลชุด Drug Effect ด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมได้ผลลัพธ์ในการประมาณค่าข้อมูลสูญหายดังแสดงในตารางที่ 22

ตารางที่ 22 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Drug Effect

จำนวนค่าสูญหาย	ตำแหน่งข้อมูลที่สูญหาย	ค่าจริง	ค่าประมาณข้อมูลสูญหาย			
			MS	CT	CL	ANN
1	y_{43}	20	30.542	38.667	35.167	21.630
2	y_{13}	16	20.500	30.667	28.667	16.809
	y_{33}	18	14.500	24.667	21.167	15.092
3	y_{12}	28	26.583	26.667	29.167	24.756
	y_{43}	20	31.083	38.667	35.167	25.159
	y_{54}	30	29.083	22.667	16.167	24.666

จากค่าประมาณค่าข้อมูลสูญหายในตารางที่ 22 การพิจารณาวิธีการประมาณค่าข้อมูลสูญหายสามารถพิจารณาได้จากการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงดังแสดงในตารางที่ 23 - 25

ตารางที่ 23 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	10.542	111.127	10.542
CopyMean Trajectory	18.667	348.444	18.667
CopyMean LOCF	15.167	230.028	15.167
Artificial Neural Network	**1.630**	**2.657**	**1.630**

จากตารางที่ 23 เมื่อพิจารณาชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียม มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

ตารางที่ 24 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	4.000	16.250	**0.500**
CopyMean Trajectory	10.667	129.778	10.667
CopyMean LOCF	7.917	85.236	7.917
Artificial Neural Network	**1.858**	**4.554**	1.049

จากตารางที่ 24 เมื่อพิจารณาชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด แต่เมื่อพิจารณาจากค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด

ตารางที่ 25 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	**4.472**	41.896	2.917
CopyMean Trajectory	9.111	134.667	3.333
CopyMean LOCF	10.056	140.917	**0.833**
Artificial Neural Network	4.579	**21.863**	1.140

จากตารางที่ 25 เมื่อพิจารณาชุดข้อมูล Drug Effect กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยแทนที่ด้วยค่าเฉลี่ยมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์น้อยที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด

1.2 ข้อมูลชุด Skydrive

ข้อมูลชุดนี้ประกอบไปด้วย k = 5, n = 11 ข้อมูลที่ได้ถูกนำมาหาค่าของสถิติพรรณาจะได้เวกเตอร์ค่าเฉลี่ย เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วม และเมทริกซ์สหสัมพันธ์ของข้อมูลชุด Skydrive แสดงได้ดังนี้

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 76.91 & 92.27 & 98.18 & 122.09 & 86.18 \end{bmatrix}'$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 49.75 & -6.77 & -9.20 & 2.04 & 13.05 \\ -6.77 & 46.64 & -12.30 & -5.02 & 27.12 \\ -9.20 & -12.30 & 58.80 & 1.79 & -23.59 \\ 2.04 & -5.02 & 1.79 & 33.64 & -3.61 \\ 13.05 & 27.12 & -23.59 & -3.61 & 48.70 \end{bmatrix}$$

$$\hat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & -0.14 & -0.17 & 0.05 & 0.27 \\ -0.14 & 1 & -0.23 & -0.13 & 0.57 \\ -0.17 & -0.23 & 1 & 0.04 & -0.44 \\ 0.05 & -0.13 & 0.04 & 1 & -0.09 \\ 0.27 & 0.57 & -0.44 & -0.09 & 1 \end{bmatrix}$$

จากการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำในข้อมูลชุด Sky Drive ด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการCopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมได้ผลลัพธ์ในการประมาณค่าข้อมูลสูญหายดังแสดงในตารางที่ 26

ตารางที่ 26 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Skydrive

จำนวนค่าสูญหาย	ตำแหน่งข้อมูลที่สูญหาย	ค่าจริง	ค่าประมาณข้อมูลสูญหาย			
			MS	CT	CL	ANN
1	y_{24}	132.82	113.107	87.103	92.405	125.183
2	y_{92}	104.71	97.033	99.665	82.581	117.892
	y_{24}	132.82	113.207	87.103	92.405	134.427
3	y_{24}	132.82	113.522	87.103	92.405	131.023
	y_{74}	112.51	111.852	85.433	104.903	98.645
	y_{65}	90.56	85.410	93.555	117.335	89.461

จากค่าประมาณค่าข้อมูลสูญหายในตารางที่ 26 การพิจารณาวิธีการประมาณค่าข้อมูลสูญหายสามารถพิจารณาได้จากการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าสัมบูรณ์ของค่าความเอนเอียงดังแสดงในตารางที่ 27 - 29

ตารางที่ 27 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	19.714	388.622	19.714
CopyMean Trajectory	45.718	2090.090	45.718
CopyMean LOCF	40.416	1633.413	40.416
Artificial Neural Network	**7.637**	**58.327**	**7.637**

จากตารางที่ 27 เมื่อพิจารณาชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

ตารางที่ 28 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	13.645	221.794	13.645
CopyMean Trajectory	25.381	1057.771	25.381
CopyMean LOCF	31.272	1061.553	31.272
Artificial Neural Network	**7.394**	**88.173**	**7.394**

จากตารางที่ 28 เมื่อพิจารณาชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

ตารางที่ 29 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	8.368	133.118	8.368
CopyMean Trajectory	25.263	944.084	23.267
CopyMean LOCF	24.933	802.729	7.083
Artificial Neural Network	**5.587**	**65.559**	**5.587**

จากตารางที่ 26 เมื่อพิจารณาชุดข้อมูล Sky Drive กรณีสุ่มค่าสูญหาย 3 ค่าจะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

1.3 ข้อมูลชุด Fecal Fat

ข้อมูลชุดนี้ประกอบไปด้วย k = 4, n = 6 ข้อมูลที่ได้ถูกนำมาหาค่าของสถิติพรรณาจะได้เวกเตอร์ค่าเฉลี่ย เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วม และเมทริกซ์สหสัมพันธ์ของข้อมูลชุด Fecal Fat แสดงได้ดังนี้

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 38.08 & 16.53 & 17.42 & 31.07 \end{bmatrix}'$$

$$\hat{\Sigma} = \begin{bmatrix} 420.92 & 164.27 & 145.17 & 386.96 \\ 164.27 & 147.87 & 137.80 & 211.11 \\ 145.17 & 137.80 & 139.48 & 218.04 \\ 386.96 & 211.11 & 218.04 & 490.62 \end{bmatrix}$$

$$\hat{\boldsymbol{\rho}} = \begin{bmatrix} 1.00 & 0.66 & 0.60 & 0.85 \\ 0.66 & 1.00 & 0.96 & 0.78 \\ 0.60 & 0.96 & 1.00 & 0.83 \\ 0.85 & 0.78 & 0.83 & 1.00 \end{bmatrix}$$

จากการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำในข้อมูลชุด Fecal Fat ด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมได้ผลลัพธ์ในการประมาณค่าข้อมูลสูญหายดังแสดงในตารางที่ 30

ตารางที่ 30 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในข้อมูลชุด Fecal Fat

จำนวนค่าสูญหาย	ตำแหน่งข้อมูลที่สูญหาย	ค่าจริง	ค่าประมาณข้อมูลสูญหาย			
			MS	CT	CL	ANN
1	y_{13}	3.4	14.070	21.400	10.825	1.925
2	y_{13}	3.4	12.613	21.400	10.825	9.264
	y_{43}	4.6	-2.188	6.600	5.100	2.034
3	y_{23}	23.1	18.258	26.467	22.367	23.100
	y_{43}	4.6	-1.608	6.600	5.100	4.751
	y_{53}	25.6	46.058	54.267	31.042	25.600

จากค่าประมาณค่าข้อมูลสูญหายในตารางที่ 30 การพิจารณาวิธีการประมาณค่าข้อมูลสูญหายสามารถพิจารณาได้จากการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าความเอนเอียงดังแสดงในตารางที่ 31 - 33

ตารางที่ 31 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	10.670	113.849	10.670
CopyMean Trajectory	18.000	324.000	18.000
CopyMean LOCF	7.425	55.131	7.425
Artificial Neural Network	**1.475**	**2.176**	**1.475**

จากตารางที่ 31 เมื่อพิจารณาชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

ตารางที่ 32 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	8.000	65.470	**1.213**
CopyMean Trajectory	10.000	164.000	10.000
CopyMean LOCF	**3.963**	27.690	3.963
Artificial Neural Network	4.215	**20.485**	1.649

จากตารางที่ 29 เมื่อพิจารณาชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วย CopyMean LOCF มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์น้อยที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด และวิธีการประมาณค่าข้อมูลสูญหายด้วยแทนที่ด้วยค่าเฉลี่ยมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด

ตารางที่ 33 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	10.503	160.176	3.136
CopyMean Trajectory	11.344	279.037	11.344
CopyMean LOCF	2.225	10.133	1.736
Artificial Neural Network	**0.050**	**0.008**	**0.050**

จากตารางที่ 33 เมื่อพิจารณาชุดข้อมูล Fecal Fat กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด

ส่วนที่ 2 ผลการวิจัยโดยใช้ชุดข้อมูลจำลอง

การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลอง ได้จำลองข้อมูลจากการแจกแจงแบบปรกติพหุซึ่งกำหนดพารามิเตอร์ดังนี้

1) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

2) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.3;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

3) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.5;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

4) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.7;\ i \neq j\ ,i, j = 1,2,3,4$$
$$k = 5, n = 4$$

5) ข้อมูลถูกสุ่มมาจากการแจกแจงปรกติพหุ 4 ตัวแปรโดยกำหนดพารามิเตอร์ได้แก่

$$\mu_i = 20;\ i = 1,2,3,4$$
$$\sigma_i^2 = 25;\ i = 1,2,3,4$$
$$\rho_{ij} = 0.9;\ i \neq j\ ,i,\ j = 1,2,3,4$$
$$k = 5, n = 4$$

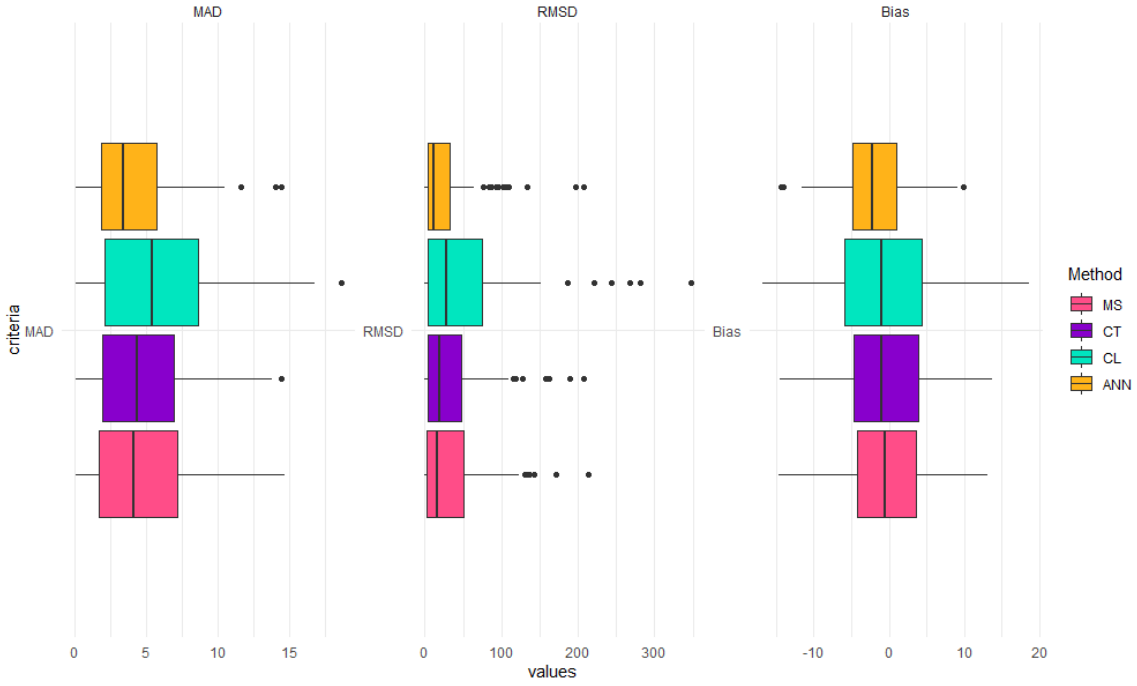
ในงานวิจัยนี้ได้จำลองข้อมูลจากโปรแกรม R เวอร์ชั่น 3.6.1 โดยทำการจำลองข้อมูลซ้ำในแต่ละพารามิเตอร์จำนวน 1000 รอบ และนำเสนอข้อมูลในรูปของตาราง และแผนภาพกล่องดังนี้

2.1 ชุดข้อมูลจำลองที่ 1 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0)

ตารางที่ 34 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	4.775	36.097	**0.288**
CopyMean Trajectory	4.764	34.689	0.458
CopyMean LOCF	5.657	49.701	0.316
Artificial Neural Network	**4.131**	**26.606**	2.033

จากตารางที่ 34 เมื่อพิจารณาชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 13



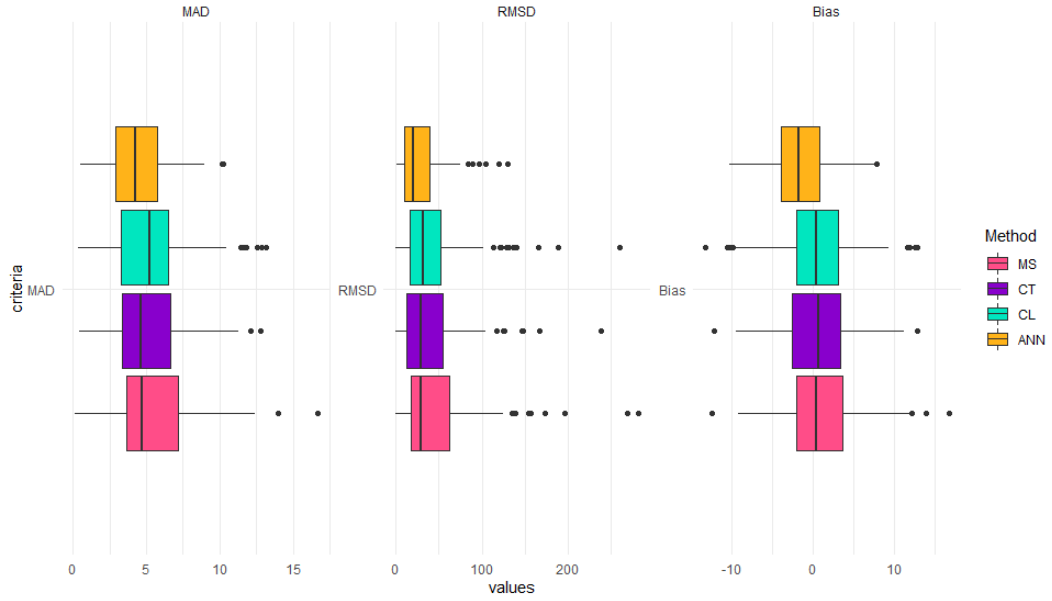
ภาพที่ 13 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 1 ค่า

เมื่อพิจารณาจากแผนภาพที่ 13 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง คือมีการกระจายของค่าประเมินน้อยและมีค่าประเมินน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น

ตารางที่ 35 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	5.535	49.038	0.941
CopyMean Trajectory	5.043	40.791	0.605
CopyMean LOCF	5.369	43.816	**0.552**
Artificial Neural Network	**4.397**	**28.685**	1.412

จากตารางที่ 35 เมื่อพิจารณาชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 14



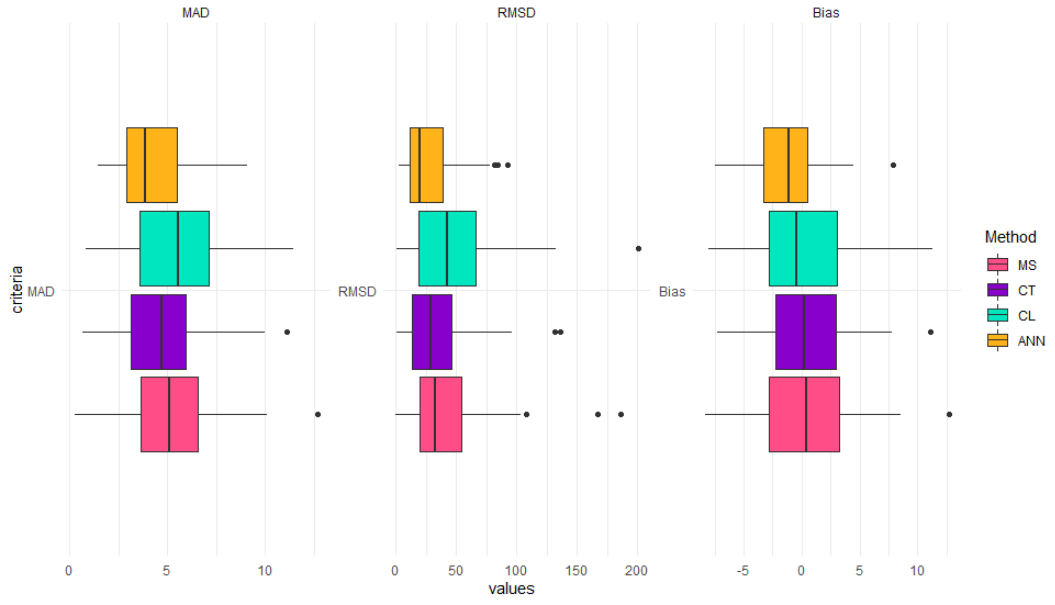
ภาพที่ 14 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 2 ค่า

เมื่อพิจารณาจากแผนภาพที่ 14 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง คือมีการกระจายของค่าประเมินน้อยและมีค่าประเมินน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น

ตารางที่ 36 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	5.153	40.221	**0.160**
CopyMean Trajectory	4.688	33.301	0.255
CopyMean LOCF	5.454	47.389	0.277
Artificial Neural Network	**4.191**	**27.328**	1.225

จากตารางที่ 36 เมื่อพิจารณาชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุดดังแสดงในภาพที่ 15



ภาพที่ 15 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 1 กรณีสุ่มค่าสูญหาย 3 ค่า

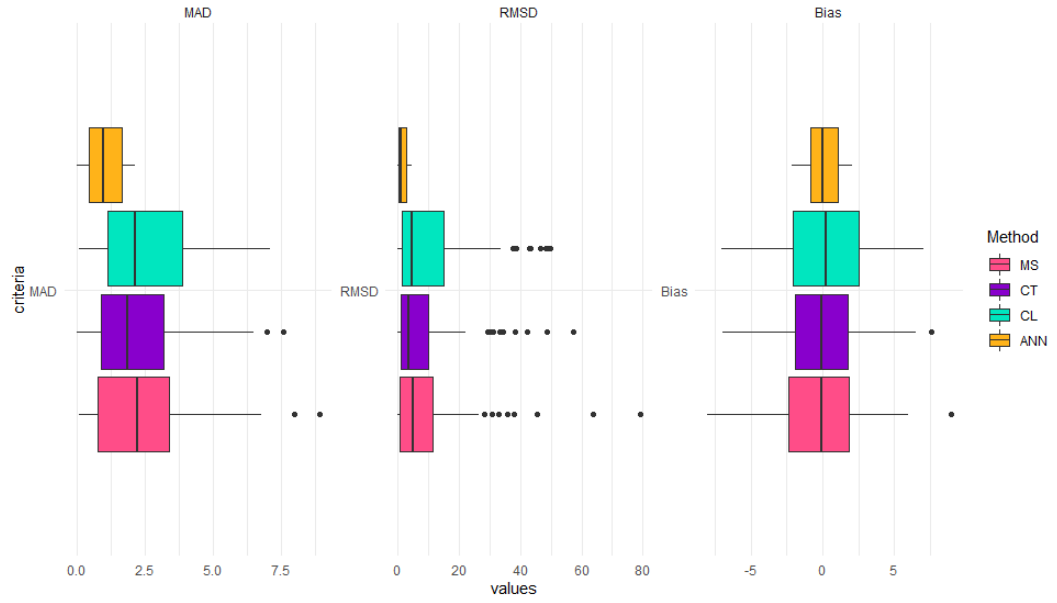
เมื่อพิจารณาจากแผนภาพที่ 15 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง คือมีการกระจายของค่าประเมินน้อยและมีค่าประเมินน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น

2.2 ชุดข้อมูลจำลองที่ 2 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.3)

ตารางที่ 37 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	2.407	9.350	0.071
CopyMean Trajectory	2.230	7.938	**0.023**
CopyMean LOCF	2.638	10.440	0.152
Artificial Neural Network	**1.030**	**1.517**	0.057

จากตารางที่ 37 เมื่อพิจารณาชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 16



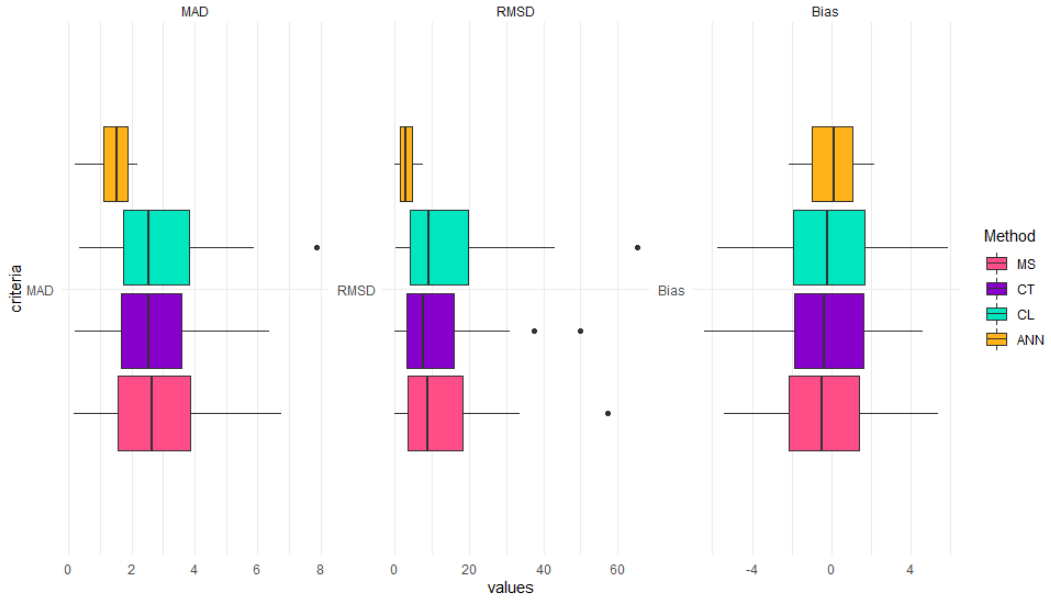
ภาพที่ 16 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า

เมื่อพิจารณาจากแผนภาพที่ 16 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น และมีการกระจายของค่าประเมินค่าข้อมูลสูญหายทั้ง 3 ค่าน้อยอีกด้วย

ตารางที่ 38 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	2.777	11.800	0.330
CopyMean Trajectory	2.645	10.737	0.273
CopyMean LOCF	2.848	12.795	0.126
Artificial Neural Network	**1.459**	**3.107**	**0.056**

จากตารางที่ 38 เมื่อพิจารณาชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด ดังแสดงในภาพที่ 17



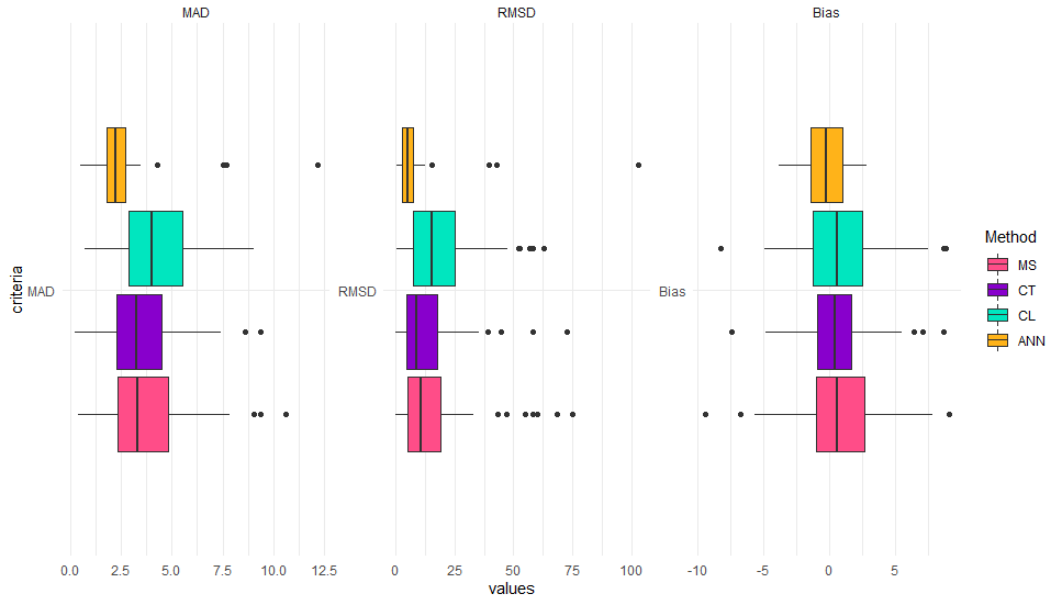
ภาพที่ 17 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 2 ค่า

เมื่อพิจารณาจากแผนภาพที่ 17 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น และมีการกระจายของค่าประเมินค่าข้อมูลสูญหายทั้ง 3 ค่าน้อยอีกด้วย

ตารางที่ 39 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	3.730	15.228	0.697
CopyMean Trajectory	3.490	13.094	0.653
CopyMean LOCF	4.230	18.269	0.784
Artificial Neural Network	**2.364**	**6.853**	**0.203**

จากตารางที่ 39 เมื่อพิจารณาชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด ดังแสดงในภาพที่ 18



ภาพที่ 18 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 2 กรณีสุ่มค่าสูญหาย 3 ค่า

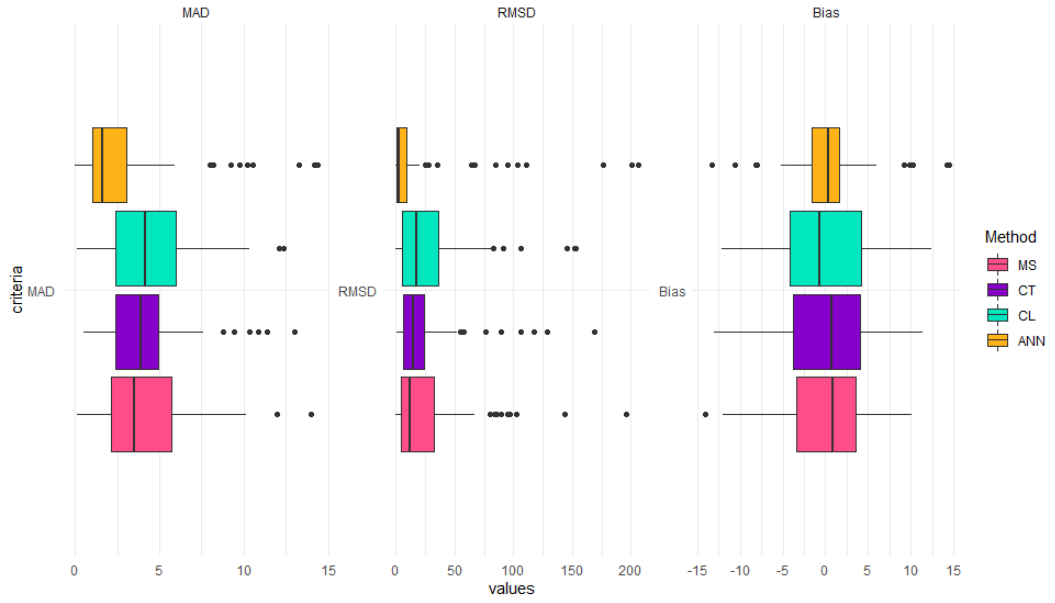
เมื่อพิจารณาจากแผนภาพที่ 18 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น และมีการกระจายของค่าประเมินค่าข้อมูลสูญหายทั้ง 3 ค่าน้อยอีกด้วย

2.3 ชุดข้อมูลจำลองที่ 3 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.5)

ตารางที่ 40 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	4.145	24.433	0.191
CopyMean Trajectory	4.054	22.346	0.132
CopyMean LOCF	4.438	27.168	**0.025**
Artificial Neural Network	**2.672**	**15.858**	0.227

จากตารางที่ 40 เมื่อพิจารณาชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนน้อยที่สุด ดังแสดงในภาพที่ 19



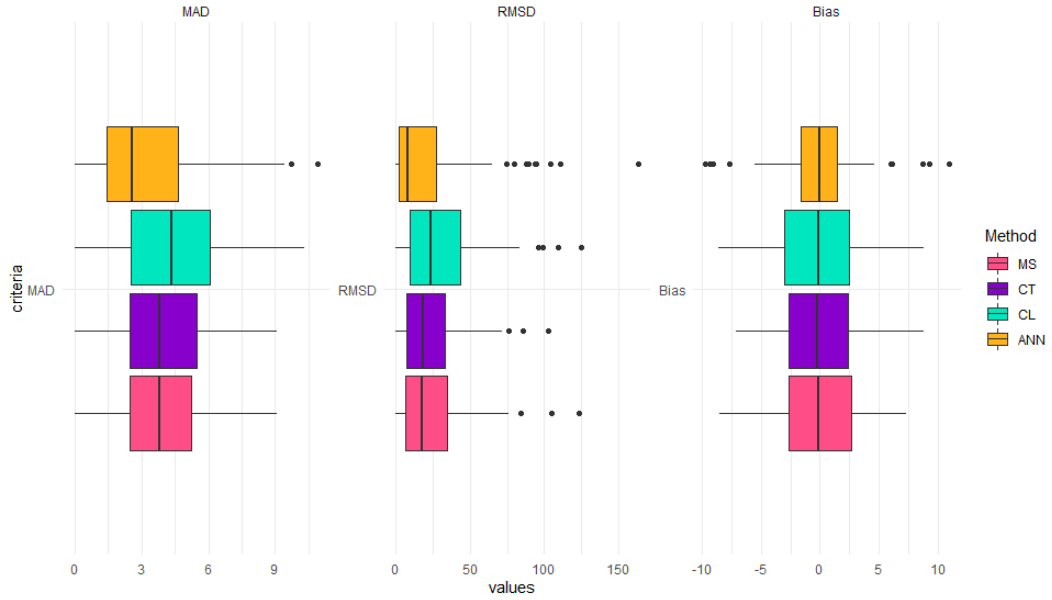
ภาพที่ 19 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 1 ค่า

เมื่อพิจารณาจากแผนภาพที่ 19 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น แต่จะเห็นว่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ค่าประเมินทั้ง 3 ค่าเกิดค่านอกกลุ่มมากที่สุด

ตารางที่ 41 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	4.023	24.320	0.236
CopyMean Trajectory	3.939	23.311	0.190
CopyMean LOCF	4.426	29.981	0.228
Artificial Neural Network	**3.224**	**21.065**	**0.129**

จากตารางที่ 41 เมื่อพิจารณาชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังนั้นในกรณีนี้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงมีประสิทธิภาพมากที่สุด ดังแสดงในภาพที่ 20



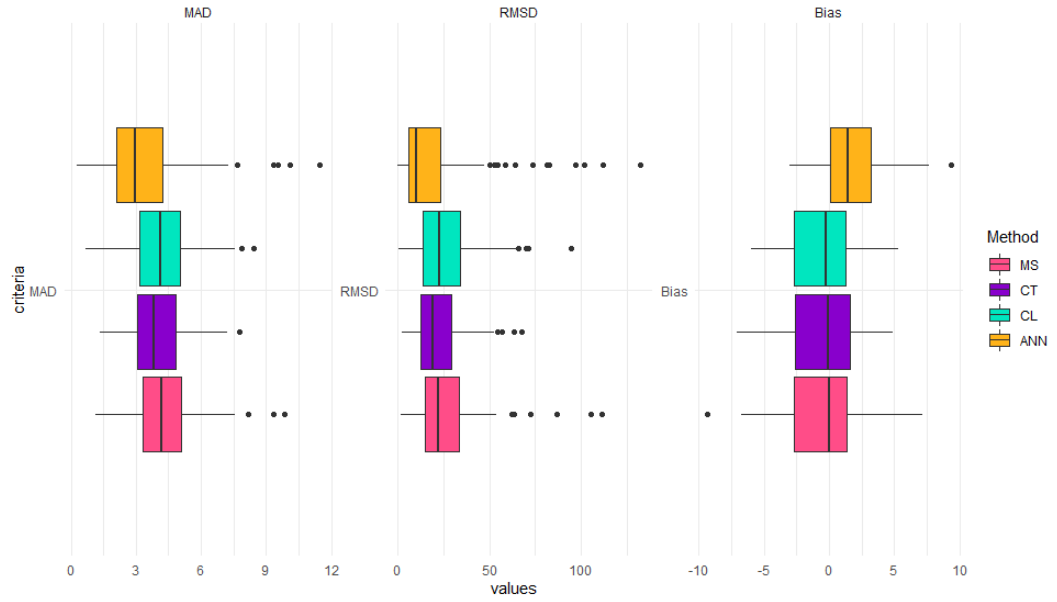
ภาพที่ 20 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาp ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 2 ค่า

เมื่อพิจารณาจากแผนภาพที่ 20 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น แต่จะเห็นว่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ค่าประเมินทั้ง 3 ค่าเกิดค่านอกกลุ่มมากที่สุด

ตารางที่ 42 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	4.288	26.499	0.493
CopyMean Trajectory	3.925	22.183	**0.491**
CopyMean LOCF	4.236	26.148	0.572
Artificial Neural Network	**3.496**	**21.288**	1.757

จากตารางที่ 42 เมื่อพิจารณาชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 21



ภาพที่ 21 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 3 กรณีสุ่มค่าสูญหาย 3 ค่า

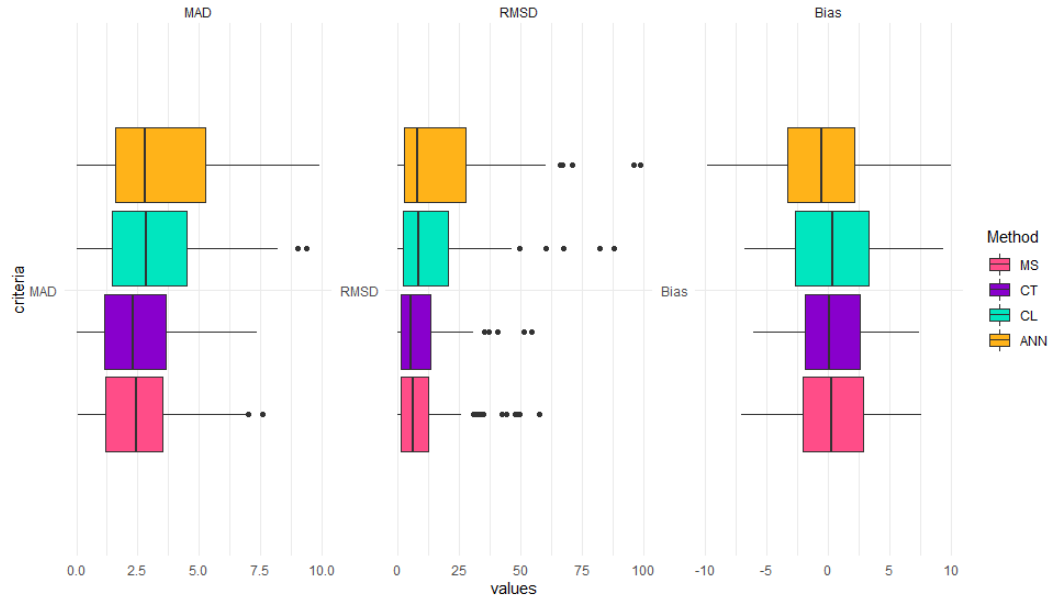
เมื่อพิจารณาจากแผนภาพที่ 21 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงน้อยกว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการอื่น แต่จะเห็นว่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ค่าประเมินทั้ง 3 ค่า เกิดค่านอกกลุ่มมากที่สุด

2.4 ชุดข้อมูลจำลองที่ 4 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.7)

ตารางที่ 43 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	2.627	10.361	**0.399**
CopyMean Trajectory	**2.583**	**10.024**	0.517
CopyMean LOCF	3.077	13.757	0.540
Artificial Neural Network	3.518	18.368	0.552

จากตารางที่ 43 เมื่อพิจารณาชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 22



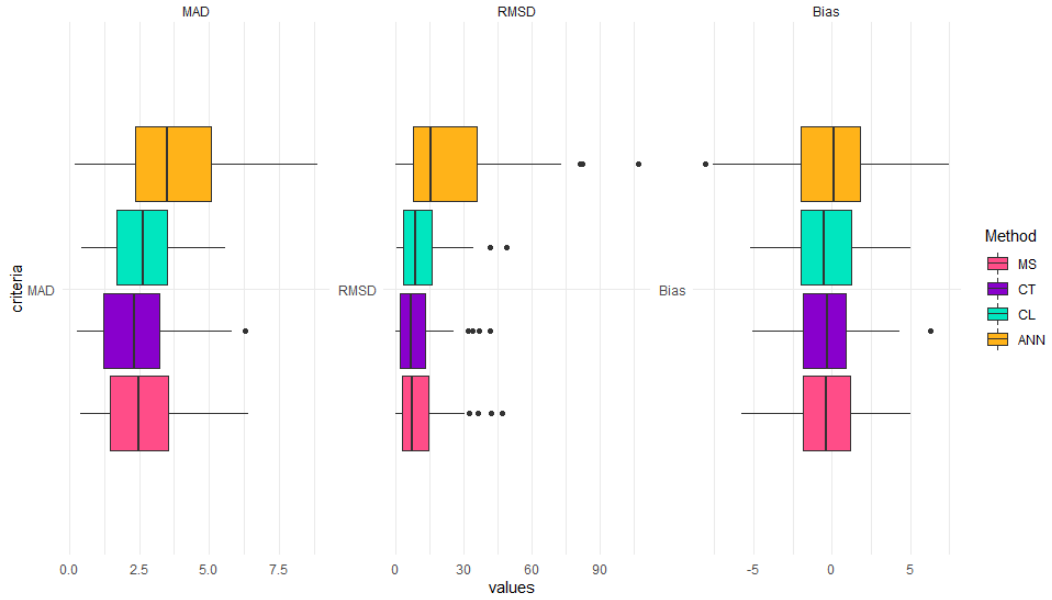
ภาพที่ 22 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 1 ค่า

เมื่อพิจารณาจากแผนภาพที่ 22 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมาก ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีแทนที่ด้วยค่าเฉลี่ย และวิธีการ CopyMean Trajectory มีค่าการประเมินและการกระจายใกล้เคียงกัน

ตารางที่ 44 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	2.623	10.399	0.221
CopyMean Trajectory	**2.411**	**9.146**	0.319
CopyMean LOCF	2.647	11.188	0.316
Artificial Neural Network	3.833	22.683	**0.056**

จากตารางที่ 44 เมื่อพิจารณาชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 23



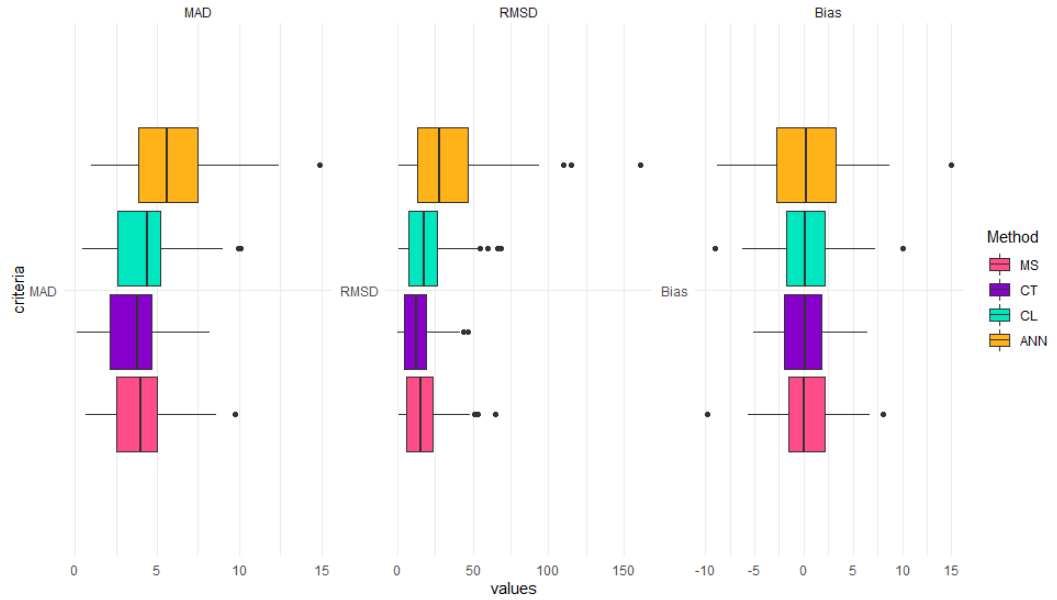
ภาพที่ 23 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 2 ค่า

เมื่อพิจารณาจากแผนภาพที่ 23 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมากมากที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีอื่นมีค่าการประเมินและการกระจายใกล้เคียงกัน

ตารางที่ 45 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	3.967	17.045	0.156
CopyMean Trajectory	**3.549**	**13.712**	0.067
CopyMean LOCF	4.259	19.498	**0.062**
Artificial Neural Network	5.701	33.921	0.380

จากตารางที่ 45 เมื่อพิจารณาชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 24



ภาพที่ 24 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 4 กรณีสุ่มค่าสูญหาย 3 ค่า

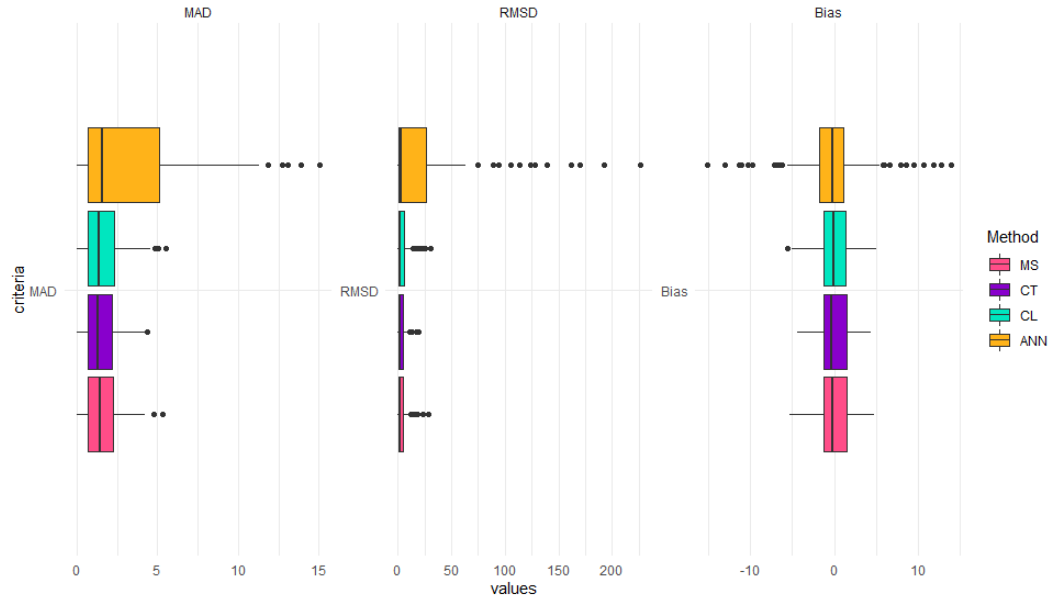
เมื่อพิจารณาจากแผนภาพที่ 24 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมากที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีอื่นมีค่าการประเมินและการกระจายใกล้เคียงกัน

2.5 ชุดข้อมูลจำลองที่ 5 (เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9)

ตารางที่ 46 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 1 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	1.621	3.926	0.059
CopyMean Trajectory	**1.524**	**3.427**	0.072
CopyMean LOCF	1.704	4.648	**0.005**
Artificial Neural Network	3.280	24.060	0.210

จากตารางที่ 46 เมื่อพิจารณาชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 1 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 19



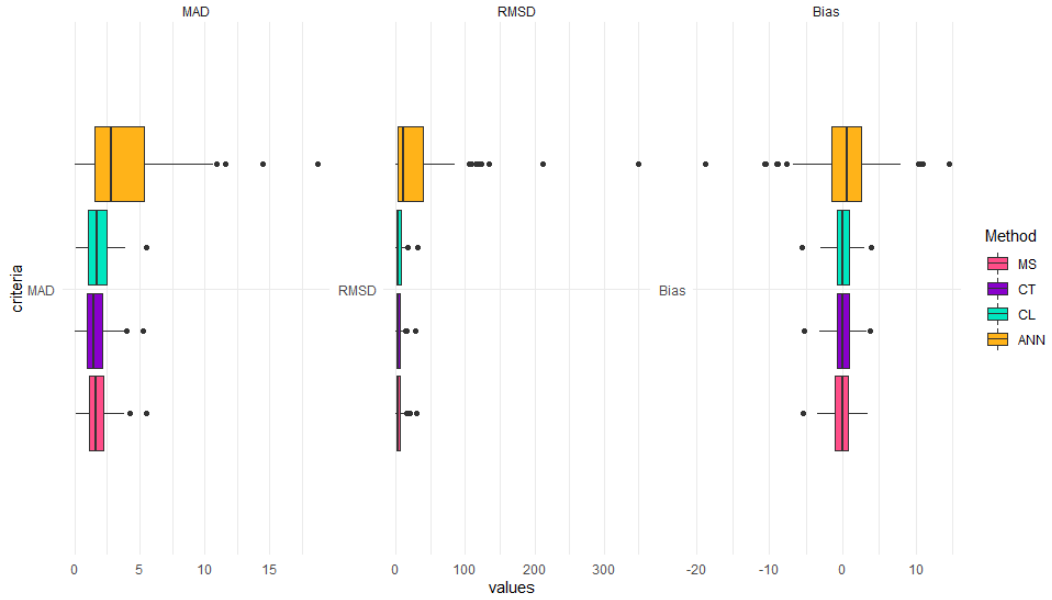
ภาพที่ 25 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 1 ค่า

เมื่อพิจารณาจากแผนภาพที่ 19 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมากและเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เกิดค่านอกกลุ่มมากที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีอื่นมีค่าการประเมินใกล้เคียงกัน

ตารางที่ 47 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 2 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	1.709	4.613	0.066
CopyMean Trajectory	**1.596**	**4.076**	0.040
CopyMean LOCF	1.780	4.872	**0.027**
Artificial Neural Network	3.969	31.147	0.331

จากตารางที่ 47 เมื่อพิจารณาชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 2 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 20



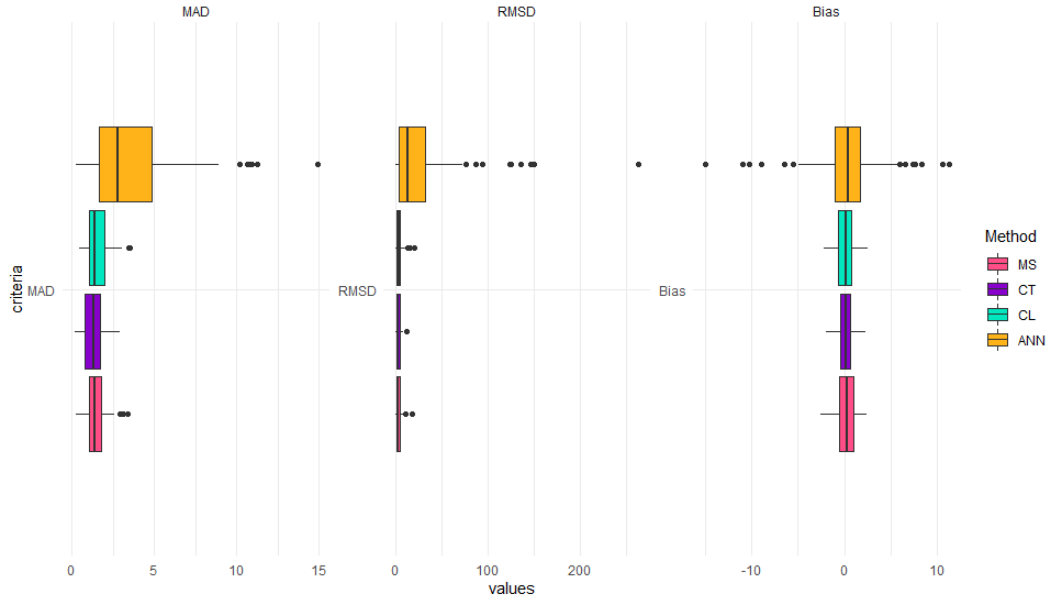
ภาพที่ 26 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย
ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 2 ค่า

เมื่อพิจารณาจากแผนภาพที่ 20 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมากและเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เกิดค่านอกกลุ่มมากที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีอื่นมีค่าการประเมินใกล้เคียงกัน

ตารางที่ 48 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 3 ค่า

วิธีการประมาณค่าข้อมูลสูญหาย	ค่าประเมินวิธีการประมาณค่าข้อมูลสูญหาย		
	MAD	RMSD	|Bias|
Mean Substitution	1.413	3.275	0.244
CopyMean Trajectory	**1.305**	**2.796**	0.165
CopyMean LOCF	1.500	3.633	**0.134**
Artificial Neural Network	3.716	27.563	0.411

จากตารางที่ 48 เมื่อพิจารณาชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 3 ค่า จะเห็นว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory มีค่าการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุด แต่เมื่อพิจารณาจากค่าสัมบูรณ์ของค่าความเอนเอียงแล้ววิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ดังแสดงในภาพที่ 21



ภาพที่ 27 แผนภาพกล่องแสดงค่าการประเมินวิธีการประมาณค่าข้อมูลสูญหาย ในชุดข้อมูลจำลองที่ 5 กรณีสุ่มค่าสูญหาย 3 ค่า

เมื่อพิจารณาจากแผนภาพที่ 27 วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีผลลัพธ์ของค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียง มีการกระจายของค่าประเมินมากและเป็นวิธีการประมาณค่าข้อมูลสูญหายที่เกิดค่านอกกลุ่มมากที่สุด ส่วนวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีอื่นมีค่าการประเมินใกล้เคียงกัน

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

1. สรุปผลการวิจัย

จากผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมโดยใช้ค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าความเอนเอียงเป็นเกณฑ์ในการประเมิน โดยแบ่งการวิจัยออกเป็นสองส่วนคือส่วนที่ 1 คือผลการวิจัยจากชุดข้อมูลจริง และส่วนที่ 2 คือผลการวิจัยจากชุดข้อมูลจำลองดังนี้

ส่วนที่ 1 สรุปผลการวิจัยโดยใช้ชุดข้อมูลจริงซึ่งในชุดข้อมูลจริงประกอบด้วยข้อมูลชุด Drug Effect ข้อมูลชุด Skydive และ ข้อมูลชุด Fecal Fat

จากผลการประมินวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมในชุดข้อมูลจริง 3 ชุดข้อมูลที่สุ่มค่าข้อมูลสูญหายแบบสูญหายอย่างสุ่มสมบูรณ์ จำนวนชุดข้อมูลละ 1 2 และ 3 ค่าตามลำดับพบว่าได้ผลลัพธ์คือเมื่อใช้ค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง และค่าสัมบูรณ์ของค่าความเอนเอียงเป็นเกณฑ์ พบว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมมีค่าประเมินน้อยที่สุดในเกือบทุกชุดข้อมูล ยกเว้นในข้อมูลชุด Drug Effect ในกรณีที่สุ่มค่าข้อมูลสูญหาย 2 ค่า วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยเป็นวิธีการที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด และในกรณีที่สุ่มค่าข้อมูลสูญหาย 3 ค่า วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ยมีค่าประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์น้อยที่สุด และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean LOCF มีค่าประเมินด้วยค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด

ดังนั้นวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมจึงเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพมากที่สุดในชุดข้อมูลจริง ดังแสดงในตารางที่ 49

ตารางที่ 49 ผลการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจริง

ชุดข้อมูล	จำนวนข้อมูลสูญหาย	วิธีการประเมินค่าข้อมูลสูญหาย		
		MAD	RMSD	|Bias|
Drug Effect	1	ANN	ANN	ANN
	2	ANN	ANN	MS
	3	MS	ANN	CL
Skydrive	1	ANN	ANN	ANN
	2	ANN	ANN	ANN
	3	ANN	ANN	ANN
Fecal Fat	1	ANN	ANN	ANN
	2	CL	ANN	MS
	3	ANN	ANN	ANN

จากผลการวิจัยทั้งในชุดข้อมูลจริงชุด Drug Effect ข้อมูลชุด Skydive และ ข้อมูลชุด Fecal Fat และในชุดข้อมูลจำลองทั้ง 5 ชุดพบว่าเมื่อสุ่มค่าข้อมูลสูญหาย 1 2 และ 3 ค่ามีผลการประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ และรากของค่าคลาดเคลื่อนกำลังสองไม่แตกต่างกันคือในชุดข้อมูลจริงส่วนใหญ่วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดี ที่สุดในการประมาณค่าข้อมูลสูญหาย

ส่วนที่ 2 สรุปผลการวิจัยโดยใช้ชุดข้อมูลจำลองซึ่งในชุดข้อมูลจำลอง ได้จำลองข้อมูลจากการแจกแจงปรกติพหุ 4 ตัวแปร (k=4) ขนาดตัวอย่างคือ 5 (n=5) โดยกำหนดพารามิเตอร์ดังนี้ เวกเตอร์ค่าเฉลี่ยคือ $\mu_i = 20; i = 1,2,3,4$ และ ความแปรปรวนคือ $\sigma_i^2 = 25; i = 1,2,3,4$ และกำหนดค่าสหสัมพันธ์แตกต่างกันในแต่ละชุดข้อมูลจำลอง คือ $\rho_{ij} = 0,\ 0.3,\ 0.5,\ 0.7,\ 0.9;$ $i \neq j\ , i, j = 1,2,3,4$ ตามลำดับ

จากผลการประมินวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean Trajectory วิธีการ CopyMean LOCF และวิธีการโครงข่ายประสาทเทียมในชุดข้อมูลจำลองในชุดข้อมูลจำลองชุดที่ 1, 2 และ 3 ซึ่งกำหนดค่าสหสัมพันธ์ได้แก่ $\rho_{ij} = 0,\ 0.3,\ 0.5;\ i \neq j$

$,i,j=1,2,3,4$ ตามลำดับ พบว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่มีค่าประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสอง น้อยที่สุดเมื่อสุ่มค่าข้อมูลสูญหายทั้ง 1, 2 และ 3 ค่า ส่วนเมื่อใช้ค่าความเอนเอียงเป็นเกณฑ์ในการประเมินวิธีการประมาณค่าข้อมูลสูญหายพบว่าในชุดข้อมูลจำลองที่ 1 เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า และ 3 ค่า วิธีการแทนที่ด้วยค่าเฉลี่ยเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ส่วนเมื่อสุ่มค่าข้อมูลสูญหาย 2 ค่า วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ในข้อมูลจำลองชุดที่ 2 เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด และเมื่อสุ่มค่าข้อมูลสูญหาย 2 ค่า และ 3 ค่า วิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ในชุดข้อมูลจำลองที่ 3 เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด เมื่อสุ่มค่าข้อมูลสูญหาย 2 ค่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุดและ เมื่อสุ่มค่าข้อมูลสูญหาย 3 ค่า วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด

ส่วนในชุดข้อมูลจำลองที่ 3 และ 4 พบว่าวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean Trajectory เป็นวิธีการที่มีค่าประเมินด้วยค่าเบี่ยงเบนสัมบูรณ์ รากของค่าคลาดเคลื่อนกำลังสองน้อยที่สุดเมื่อสุ่มค่าข้อมูลสูญหายทั้ง 1, 2 และ 3 ค่า ส่วนเมื่อใช้ค่าความเอนเอียงเป็นเกณฑ์ในการประเมิน ในชุดข้อมูลจำลองที่ 4 เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า วิธีการแทนที่ด้วยค่าเฉลี่ยเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด เมื่อสุ่มค่าข้อมูลสูญหาย 2 ค่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด และเมื่อสุ่มค่าข้อมูลสูญหาย 3 ค่า วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุด ส่วนในชุดข้อมูลจำลองที่ 5 วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีค่าสัมบูรณ์ของค่าความเอนเอียงน้อยที่สุดทั้งในกรณีที่สุ่มค่าข้อมูลสูญหาย 1, 2 และ 3 ค่า ดังแสดงในตารางที่ 50

ตารางที่ 50 ผลการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลอง

ชุดข้อมูล	จำนวนข้อมูลสูญหาย	วิธีการประเมินค่าข้อมูลสูญหาย		
		MAD	RMSD	|Bias|
dataset 1 $(\rho = 0)$	1	ANN	ANN	MS
	2	ANN	ANN	CL
	3	ANN	ANN	MS
dataset 2 $(\rho = 0.3)$	1	ANN	ANN	CT
	2	ANN	ANN	ANN
	3	ANN	ANN	ANN
dataset 3 $(\rho = 0.5)$	1	ANN	ANN	CL
	2	ANN	ANN	ANN
	3	ANN	ANN	CT
dataset 4 $(\rho = 0.7)$	1	CT	CT	MS
	2	CT	CT	ANN
	3	CT	CT	CL
dataset 5 $(\rho = 0.9)$	1	CT	CT	CL
	2	CT	CT	CL
	3	CT	CT	CL

ในชุดข้อมูลจำลองชุดที่ 1, 2 และ 3 ที่มีค่าสหสัมพันธ์น้อยคือ 0, 0.3 และ 0.5 ตามลำดับ วิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายเช่นกัน แต่ในชุดข้อมูลจำลองที่ 4 และ 5 ที่มีค่าสหสัมพันธ์มาก ได้แก่ 0.7 และ 0.9 วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุด และเมื่อพิจารณาจากชุดข้อมูลจริง จะพบว่าข้อมูลชุด Sky Drive ซึ่งเป็นชุดข้อมูลที่มีค่าสหสัมพันธ์น้อย วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายในทุกกรณี ซึ่งสอดคล้องกับผลลัพธ์ในชุดข้อมูลจำลอง

เมื่อพิจารณาค่าประเมินเมื่อกำหนดค่าข้อมูลสูญหายจำนวน 1 2 และ 3 ค่าตามลำดับพบว่าค่าประเมินด้วยเกณฑ์การประเมินทั้ง 3 วิธีมีค่าใกล้เคียงกันเมื่อพิจารณาค่าประเมินในกรณีที่สุ่มค่าข้อมูลสูญหายต่างกันจากชุดข้อมูลเดียวกัน แต่เมื่อพิจารณาค่าประเมินในกรณีที่ค่าสหสัมพันธ์ต่างกันพบว่าเมื่อค่าสหสัมพันธ์มีค่ามากขึ้นแล้วค่าประเมินด้วยเกณฑ์การประเมินทั้ง 3 วิธีมีแนวโน้มที่ลดลงเมื่อค่าสหสัมพันธ์เพิ่มขึ้น

2. อภิปรายผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองเมื่อประมาณค่าข้อมูลสูญหายโดยใช้วิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean และวิธีโครงข่ายประสาทเทียม และเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายโดยใช้ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าความเอนเอียงเป็นเกณฑ์ในการประเมินประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลองทั้ง 5 ชุดผลการทดลองในชุดข้อมูลจริง และชุดข้อมูลจำลองที่ 1 – 3 (กรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์คือ 0, 0.3, 0.5) ผลการทดลองคือ วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดซึ่งสอดคล้องกับในงานวิจัยของ Gupta และ Lam

ในงานวิจัยของ Genolini และคณะ วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูล สูญหายที่เหมาะสมที่สุดในการศึกษาข้อมูลตามคาบเวลา ในงานวิจัยนี้เมื่อพิจารณาในชุดข้อมูลจำลองที่ 4 – 5 (กรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์คือ 0.7 และ 0.9) วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุด ซึ่งทั้งในงานวิจัยนี้ในกรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์มาก และในงานวิจัยของ Genolini และคณะวิธีการ CopyMean เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดเช่นเดียวกันแต่แตกต่างกันที่วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตามคาบเวลา ซึ่งอาจเกิดจากในงานวิจัยนี้ได้กำหนดค่าความแปรปรวนของทุกตัวแปรมีค่าเท่ากัน และขนาดตัวอย่างในงานวิจัยนี้มีขนาดเล็กกว่าในงานวิจัยของ Genolini และคณะในการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean ที่เลือกใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตามคาบเวลาด้วยวิธีการ Trajectory mean จึงเหมาะสมกว่าวิธีการ LOCF

3. ข้อเสนอแนะ

3.1 ในงานวิจัยครั้งนี้ได้กำหนดความแปรปรวนเท่ากันในทุกตัวแปรในการศึกษาครั้งต่อไปควรศึกษาในกรณีที่กำหนดพารามิเตอร์ความแปรปรวนไม่เท่ากันเพิ่มเติม

3.2 ในงานวิจัยครั้งนี้ได้กำหนดขนาดตัวอย่าง n = 5 ในการศึกษาครั้งต่อไปควรศึกษาในขนาดตัวอย่างอื่นเพิ่มเติม

3.3 ในงานวิจัยครั้งนี้ได้สุ่มข้อมูลจากการแจกแจงปรกติพหุ ในการศึกษาครั้งต่อไปควรศึกษาในกรณีที่สุ่มข้อมูลจากการแจกแจงอื่นเพิ่มเติม

รายการอ้างอิง

Berman, H. (Producer). (2019, September 16). Bias. *Stat Trek Teach yourself statistics*. Retrieved from https://stattrek.com/statistics/dictionary.aspx?definition=bias

Bingham, C. R., Stemmler, M., Peterson, A. C., & Graber, J. A. (1998). Imputing Missing Data Values in Repeated Measurement Within-Subjects Designs. *Methods of Psychological Research Online, 3*.

Caruana, E. J., Roman, M., Hernández, J. S., & Solli, P. (2015). Longitudinal studies. *Journal of Thoracic Disease*, E537–E540.

A comprehensive guide to repeated measures ANOVA test. (2019, june 12). *Medium*. Retrieved from https://medium.com/@dissertationserviceuk/a-comprehensive-guide-to-repeated-measures-anova-test-2b06dbfe7de8

Field, A. (2000). One Way Repeated Measures ANOVA by Hand. *discoveringstatistics*. Retrieved from https://www.discoveringstatistics.com/repository/onewayrmhand.pdf

Fritsch, S., Guenther, F., & Wright, M. N. (Producer). (2019, Febuary 7). Package 'neuralnet'. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

Genolini, C., Lacombe, A., cochard, R., & Subtil, F. (2016). CopyMean: A new method to predict monotone missing values in longitudinal studies. *computer methods and programs in biomedicine, 132*, 29–44.

Gupta, A., & Lam, M. S. (1996). Estimating Missing Values Using Neural Networks. *The Journal of the Operational Research Society, 41*, 229-238.

Kang, H. (2013). The prevention and handling of the missing data. *the Korean Society of Anesthesiologists, 64(5)*, 402-406.

Kızrak, A. (Producer). (2019, may 9). Comparison of Activation Functions for Deep Neural Networks. *towardsdatascience*. Retrieved from

https://towardsdatascience.com/comparison-of-activation-functions-for-deep-neural-networks-706ac4284c8a

Lani , J. (Producer). (2019, May 6). Handling Missing Data: Listwise Versus Pairwise Deletion. *statistics solutions*. Retrieved from https://www.statisticssolutions.com/handling-missing-data-listwise-versus-pairwise-deletion/

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. United States ofAmerica: Library of Congress Cataloging.

LUND, A., & LUND, M. (Producer). (2018a, December 21). Measures of Spread. *lærd statistics*. Retrieved from https://statistics.laerd.com/statistical-guides/measures-of-spread-absolute-deviation-variance.php

LUND, A., & LUND, M. (Producer). (2018b, Decamber 12). Repeated Measures ANOVA. *lærd statistics*. Retrieved from https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

Markgraf, B. (Producer). (2018, March 23). How to Calculate Bias. *SCIENCING*. Retrieved from https://sciencing.com/how-to-calculate-bias-13710241.html

Nielson, M. (Producer). (2019, June 2). Neural network and deep learning. *cognitivemedium.com*. Retrieved from http://neuralnetworksanddeeplearning.com

Price, V. H., & Menefee, E. (1990). Quantitative Estimation of Hair Growth I. Androgenetic Alopecia in woman : Effect of Minoxidil. *The Journal of Investigative Dermatology, 95*, 683-687.

Rubin, L. H., Witkiewitz, K., Andre, J. S., & Reilly, S. (2007). Method for handling missing data in the behavioral neurosciences:Don't throw the baby rat out with the bath water. *The journal of undergraduate neuroscience education*, A71-A77.

Shamdasani, S. (Producer). (2017, October 4). Build a Neural Network. *enlight*. Retrieved from https://enlight.nyc/projects/neural-network/

Shuttleworth, M. (2009, 26 November). Repeated Measures Design.

Singley, K. I., Hale, B. D., & Russell, D. (2012). Heart Rate, Anxiety, and Hardiness in Novice (Tandem) and Experienced (Solo) Skydivers. *Journal of Sport Behavior, 35*, 453-469.

Stephanie (Producer). (2016, October 25). RMSE: Root Mean Square Error. *Statistics How To*. Retrieved from https://www.statisticshowto.datasciencecentral.com/rmse/

Vermeulen, K. M., Post, W. J., Span, M. M., Bij, W. v. d., Koëter, G. H., & TenVergert, E. M. (2005, September 8). Incomplete quality of life data in lung transplant research: comparing cross sectional, repeated measures ANOVA, and multi-level analysis. *BMC Part of Springer Nature*.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression Methods in Biostatistics*. New York, USA: Springer Science+Business Media.

Winer, B. J. (1962). *STATISTICAL PRINCIPLES IN EXPERIMENTAL DESIGN*: McGRAW-HILL BOOK.

Zaiontz, C. (Producer). (2014, December 12). Assumptions for Statistical Tests. *Real Statistics Using Excel*. Retrieved from http://www.real-statistics.com/descriptive-statistics/assumptions-statistical-test/

ภาคผนวก

โปรแกรมคอมพิวเตอร์ที่ใช้ในงานวิจัย

```
#function for calculate y bar i dot
mi <- function(z){
  f <- rep(0,nrow(z))
  for(i in 1:nrow(z)) {
    f[i] <- mean(t(z)[,i],na.rm = T)
  }
  return(f)
}

#function for calculate y bar dot j
mj <- function(z){
  f <- rep(0,ncol(z))
  for(i in 1:ncol(z)) {
    f[i] <- mean(z[,i],na.rm = T)
  }
  return(f)
}

#function for random missing point
misspoint <- function(d,m){
  M <- t(d)
  a <- sample((ncol(M)+1):((nrow(M)-1)*ncol(M)),m,FALSE)
  return(a)
}
miss <- function(d,m,a){
  M <- t(d)
  for (i in 1:m) {
    if(a[i]%%ncol(M) != 0){
      M[ceiling(a[i]/ncol(M)), a[i]%%ncol(M)] <- NA
    }else{
```

```
      M[ceiling(a[i]/ncol(M)), ncol(M)] <- NA
    }
  }
  return(t(M))
}

#Mean Substitution
ms <- function(z,y){
  z <- t(z)
  I <- rep(0,nrow(z))
  A <- matrix(rep(NA,nrow(z)*ncol(z)),nrow = nrow(z),ncol = ncol(z))
  for(i in 1:nrow(z)) {
    for(j in 1:ncol(z)) {
      if(is.na(z[i,j])==FALSE ){
        A[i,j] <- y[j]-z[i,j]
      }
    }
  }
  for(i in 1:nrow(z)) {
    I[i] <- mean(A[i,],na.rm = T)
  }
  for(i in 1:nrow(z)) {
    for(j in 1:ncol(z)) {
      if(is.na(z[i,j])==TRUE){
        z[i,j] <- y[j]-I[i]
      }
    }
  }
  return(t(z))
}
```

```
# longitudinal imputation > LOCF method
locf <- function(M){
  L <- M
  for (i in 1:nrow(M)) {
    for (j in 1:ncol(M)) {
      if(is.na(L[i,j])==TRUE){
        L[i,j] <- L[i,j-1]
      }else{
        L[i,j] <- L[i,j]
      }
    }
  }
  return(L)
}

#CopyMean function
CM <- function(M,L){
  A <- M
  for (i in 1:nrow(A)) {
    for (j in 1:ncol(A)) {
      if(is.na(A[i,j])==TRUE){
        mi <- mean(M[i,],na.rm = TRUE)
        mLi <- mean(L[i,])
        AV <- mi-mLi
        A[i,j] <- L[i,j]+AV
      }else{
        A[i,j] <- M[i,j]
      }
    }
  }
  return(A)
```

```
}

#Trajectory Mean Imputation
TM <- function(M,I){
  L <- M
  for (i in 1:nrow(M)) {
    for (j in 1:ncol(M)) {
      if(is.na(L[i,j])==TRUE){
        L[i,j] <- I[i]
      }else{
        L[i,j] <- L[i,j]
      }
    }
  }
  return(L)
}

# Custom activation function(switch)
swt <- function(x) (2*x)/(1+exp(-x))
# เรียกใช้ฟังก์ชั่น neural network
library(neuralnet)

#remove missing point
rem <- function(M,a){
  A <- data.frame(rep(NA,(length(a))*(nrow(M)-length(a))),(nrow(M)-length(a)),(ncol(M)))
  for (j in 1: length(a)) {
    if(a[j]%%nrow(M) == 0){
      a[j] <- nrow(M)
    }else{
      a[j] <- a[j]%%nrow(M)
    }
```

```
  }
  for (i in 1:length(a)) {
    A <- M[-a,]
  }
  return(as.data.frame(A))
}
#function MAD
MAD <- function(M,R,n){
  mad <- 0
  for (i in 1:nrow(M)) {
    for (j in 1:ncol(M)) {
      mad <- mad+abs(M[i,j]-R[i,j])
    }
  }
  return(as.vector(mad/n))
}

# function RMSD
RMSD <- function(M,R,n){
  rmsd <- 0
  for (i in 1:nrow(M)) {
    for (j in 1:ncol(M)) {
      rmsd <- rmsd + (M[i,j]-R[i,j])^2
    }
  }
  return(as.vector(rmsd/n))
}

#Function Bias
Bias <- function(M,R,n){
  bias <- 0
```

```
  for (i in 1:nrow(M)) {
    for (j in 1:ncol(M)) {
      bias <- bias+(M[i,j]-R[i,j])
    }
  }
  return(as.vector(bias/n))
}

#นำเข้าชุดข้อมูลDrugEffect
V1<- c(30,14,24,38,26)
V2<- c(28,18,20,34,28)
V3<- c(16,10,18,20,14)
V4<- c(34,22,30,44,30)
df <- data.frame(V1,V2,V3,V4)

#เรียกใช้ฟังก์ชั่น mgcv
library(mgcv, lib.loc = "C:/Program Files/R/R-3.6.1/library")
library(mgcv, lib.loc = "C:/Program Files/R/R-3.6.1/library")

#กรณีสุ่มค่าข้อูลสูญหาย 1 ค่า
#สร้างเมทริกซ์ว่างสำหรับเก็บค่าผลลัพธ์
df0 <- matrix(rep(0,4*5),5,4)
MPdf1 <- rep(0,1)
SMdf1 <- df0
MSdf1 <- df0
CLdf1 <- df0
CTdf1 <- df0
nnerrordf1 <- rep(0,1)
Yhatdf1 <- rep(0,1)
ANNdf1 <- df0
MADdf1 <- rep(0,4)
```

```
RMSDdf1 <- rep(0,4)
Biasdf1 <- rep(0,4)

#สุ่มตำแหน่งของข้อมูลสูญหาย
set.seed(99)
mp <- misspoint(df,1)
MPdf1 <- mp
sm <- miss(df,1,mp)
SMdf1 <- sm
MI <- mi(sm)
MJ <- mj(sm)

#Mean Substitution imputation
MS <- ms(sm,MI)
MSdf1 <- MS

# CopyMean LOCF Imputation
lm <- locf(sm)
cl <- CM(sm,lm)
CLdf1 <- cl

# CopyMean Trajectory Imputation
tm <- TM(sm,MI)
ct <- CM(sm,tm)
CTdf1 <- ct

#ANN
rm <- rem(sm,mp)
for (j in 1:1) {
  if(mp[j]%%nrow(sm) == 0){
    smn <- as.data.frame(sm)
```

```
    pd <-smn[nrow(sm),-ceiling(mp[j]/5)]
  }else{
    smn <- as.data.frame(sm)
    pd <- smn[(mp[j]%%nrow(sm)),-ceiling(mp[j]/5)]
  }

    if(ceiling(mp[j]/5) == 1){
      nn <- neuralnet(V1~V2+V3+V4,data = rm, hidden = c(2,2),lifesign = 'none',
algorithm='backprop',act.fct = swt,  threshold = 1000 ,
                      linear.output=TRUE,learningrate = 0.0001,err.fct = 'sse')
    }else{
      if(ceiling(mp[j]/5) == 2){
        nn <- neuralnet(V2~V1+V3+V4,data = rm, hidden = c(2,2),lifesign = 'none',
algorithm='backprop',act.fct = swt,  threshold = 1000 ,
                        linear.output=TRUE,learningrate = 0.0001,err.fct = 'sse')
      }else{
        if(ceiling(mp[j]/5) == 3){
          nn <- neuralnet(V3~V1+V2+V4,data = rm, hidden = c(2,2),lifesign = 'none',
algorithm='backprop',act.fct = swt,  threshold = 1000 ,
                        linear.output=TRUE,learningrate = 0.0001,err.fct = 'sse')
        }else{
          nn <- neuralnet(V4~V1+V2+V3,data = rm, hidden = c(2,2),lifesign = 'none',
algorithm='backprop',act.fct = swt,  threshold = 1000 ,
                        linear.output=TRUE,learningrate = 0.0001,err.fct = 'sse')
        }
      }
    }
    if(as.vector(nn$result.matrix[1,1])<5){
      pred <- compute(nn,pd)
      nnet <- nn
      yhat <- pred$net.result
```

```
    }
  }
  nnerrordf1[j] <- nnet$result.matrix[1,1]
  Yhatdf1[j] <- yhat
  if(MPdf1[j]%%nrow(df0) != 0){
    ann[MPdf1[j]%%nrow(df0), ceiling(MPdf1[j]/nrow(df0))] <- yhat
  }else{
    ann[nrow(df0),ceiling(MPdf1[j]/nrow(df0))] <- yhat
  }
ANNdf1 <- ann



#Criteria Calculation
MADdf1[1] <- MAD(MS,df,1)
MADdf1[2] <- MAD(ct,df,1)
MADdf1[3] <- MAD(cl,df,1)
MADdf1[4] <- MAD(ann,df,1)
RMSDdf1[1] <- RMSD(MS,df,1)
RMSDdf1[2] <- RMSD(ct,df,1)
RMSDdf1[3] <- RMSD(cl,df,1)
RMSDdf1[4] <- RMSD(ann,df,1)
Biasdf1[1] <- Bias(MS,df,1)
Biasdf1[2] <- Bias(ct,df,1)
Biasdf1[3] <- Bias(cl,df,1)
Biasdf1[4] <- Bias(ann,df,1)
MADdf1
RMSDdf1
Biasdf1

#จำลองข้อมูลชุดที่ 1 กำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0
mu1 <- rep(20,4)
```

```
r1 <- diag(rep(1,4))
c1 <- diag(rep(5,4))
cov1 <- c1%*%r1%*%c1
for (a in 1:1000) {
  set.seed(11*a)
          mvn <- rmvn(5,mu1,cov1)
                  D11[,,a] <- mvn
                  }
```

ประวัติผู้เขียน

ชื่อ-สกุล	นลัทพร รูปหมอก
วัน เดือน ปี เกิด	1 พฤษภาคม 2538
สถานที่เกิด	สุพรรณบุรี
วุฒิการศึกษา	มหาวิทยาลัยศิลปากร
ที่อยู่ปัจจุบัน	บ้านเลขที่ 4 หมู่ที่ 4 ตำบล สระยายโสม อำเภอ อู่ทอง จังหวัด สุพรรณบุรี 72220