บทที่

1

การดำเนินการกับข้อมูลขาดหาย

Dealing with Missing Data

วุฒิ สุขเจริญ
Wut Sookcharoen

บทที่ 1

การดำเนินการกับข้อมูลขาดหาย

Dealing with Missing Data

วุฒิ สุขเจริญ [1]
Wut Sookcharoen

บทคัดย่อ

การวิจัยเกือบทุกประเภทมีโอกาสเกิดข้อมูลขาดหาย ซึ่งเกิดจากสาเหตุหลากหลาย ปัญหาของข้อมูลขาดหายทำให้การวิเคราะห์ข้อมูลขาดประสิทธิภาพและมีความยุ่งยากในการวิเคราะห์ข้อมูล เนื่องจากข้อมูลมีความลำเอียงจากข้อมูลที่ขาดหาย จึงได้พัฒนาวิธีต่าง ๆ เพื่อพยายามทดแทนข้อมูลขาดหาย การดำเนินการกับข้อมูลที่ขาดหายโดยใช้วิธี listwise deletion, pairwise deletion และการแทนค่าด้วยค่าเฉลี่ยเลขคณิตหรือค่าฐานนิยม เป็นวิธีที่มีความถูกต้องต่ำ ผู้เขียนแนะนำให้ใช้การดำเนินการกับข้อมูลขาดหายโดยใช้วิธีที่มีความถูกต้องสูง เช่น full information maximum-likelihood, expectation-maximization algorithm และ multiple Imputation

คำสำคัญ : ข้อมูลขาดหาย การดำเนินการ

Abstract

In almost all research, there is the potential for missing data. Missing data can occur for various reasons. Problems associated with missing values are loss of efficiency, complications in handling and analyzing the data, and bias resulting from differences between missing data and complete data. A variety of methods have been developed to

[1] อาจารย์ประจำคณะบริหารธุรกิจ สถาบันเทคโนโลยีไทย-ญี่ปุ่น, ผู้ช่วยศาสตราจารย์ (Ph.D.)

attempt to compensate for missing data. The author suggests that listwise deletion, pairwise deletion and mean/mode substitution are poor methods for handling missing data, whereas full information maximum-likelihood, expectation-maximization algorithm, and multiple imputations are recommended alternatives to this approach.

Key words : Missing Data ; Dealing

บทนำ

ข้อมูลขาดหาย คือ ข้อมูลที่ไม่ครบถ้วนหรือไม่สมบูรณ์ สามารถเกิดขึ้นได้กับงานวิจัยทุกประเภท โดยเฉพาะอย่างยิ่งงานวิจัยเชิงปริมาณที่ใช้แบบสอบถามเป็นเครื่องมือในการเก็บรวบรวมข้อมูล

Chao-Ying Joanne Peng, et al. (2006) ได้ศึกษาผลงานวิจัยตีพิมพ์ในวารสารด้านการศึกษาและจิตวิทยาจำนวน 11 ฉบับ ระหว่างปี ค.ศ.1998-2004 พบว่า ร้อยละ 48 ของผลงานที่ตีพิมพ์ระบุว่าการเก็บรวบรวมข้อมูลมีข้อมูลขาดหาย

Angela M Wood, Ian R White and Simon G Thompson (2004) ได้ศึกษาผลการวิจัยโดยการสุ่มตัวอย่างจากผลงานวิจัยที่ได้รับการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet and New England Journal of Medicine ระหว่างเดือนกรกฎาคมถึงธันวาคม จำนวน 71 ชิ้น พบว่า มีผลงานวิจัยจำนวน 63 ชิ้น คิดเป็นร้อยละ 89 ระบุว่ามีข้อมูลขาดหายบางส่วน และมีผลงานวิจัยจำนวน 13 ชิ้น คิดเป็นร้อยละ 18 ที่ระบุว่ามีข้อมูลขาดหายเกินกว่าร้อยละ 20 ของกลุ่มตัวอย่าง จะเห็นได้ว่าข้อมูลขาดหายพบได้ปกติในการทำวิจัย

ถึงแม้ข้อมูลขาดหายจะพบได้ปกติในการเก็บรวบรวมข้อมูลเพื่อการวิจัย แต่ยังไม่มีข้อสรุปที่ชัดเจนในแนวทางการดำเนินการเกี่ยวกับข้อมูลขาดหาย Yiran Dong and Chao-Ying Joanne Peng (2013) ได้ศึกษาการดำเนินการเกี่ยวกับข้อมูลขาดหาย พบว่างานวิจัยที่มีข้อมูลขาดหายร้อยละ 37 ไม่ได้มีการดำเนินการใด ๆ เกี่ยวกับข้อมูลขาดหาย

การดำเนินการกับข้อมูลขาดหายมักเป็นดุลยพินิจของนักวิจัย ซึ่งอาจทำให้ผลการวิจัยมีความคลาดเคลื่อนหรือมีความลำเอียง ดังนั้นนักวิจัยจึงจำเป็นต้องมีแนวทาง

ในการป้องกันเพื่อลดการเกิดข้อมูลขาดหาย และต้องมีความรู้เกี่ยวกับการดำเนินการเกี่ยวกับข้อมูลหายขาด เพื่อเลือกใช้วิธีการดำเนินการกับข้อมูลหายขาดอย่างเหมาะสม

สาเหตุและแนวทางการป้องกันการเกิดข้อมูลขาดหาย

ข้อมูลขาดหายเกิดได้จากหลายสาเหตุ แต่เมื่อพิจารณาจากแหล่งที่มาของการเกิดข้อมูลขาดหาย สามารถแบ่งการเกิดข้อมูลขาดหายออกเป็น 2 กลุ่ม ดังนี้

1. การเกิดข้อมูลขาดหายจากกระบวนการวิจัย ได้แก่

- **การกำหนดประเด็นการวิจัย** การดำเนินการวิจัยมีต้นทุนด้านการใช้ทรัพยากร บางครั้งนักวิจัยได้กำหนดประเด็นที่ต้องการศึกษาไว้หลากหลายเกินไป ทำให้ผู้ให้ข้อมูลไม่สามารถให้ข้อมูลได้ครบทุกประเด็น ส่งผลให้เกิดข้อมูลขาดหาย เนื่องจากผู้ให้ข้อมูลไม่สามารถให้ข้อมูลในบางประเด็น เพื่อป้องกันปัญหาดังกล่าว นักวิจัยควรกำหนดประเด็นการศึกษาให้ตรงประเด็นและกระชับ

- **การออกแบบแบบสอบถามที่ใช้เวลาตอบมากเกินไป** ระยะเวลารวมในการทำแบบสอบถามเป็นปัจจัยหนึ่งที่เป็นสาเหตุทำให้เกิดข้อมูลขาดหาย แบบสอบถามที่ใช้เวลาตอบมากเกินไป ผู้ตอบแบบสอบถามอาจตอบไม่ครบคำถามข้อท้าย ๆ ของแบบสอบถาม ผู้ตอบแบบสอบถามอาจเร่งรีบตอบ ทำให้นักวิจัยการตลาดได้ข้อมูลที่ไม่ถูกต้อง จากการสำรวจความคิดเห็นของผู้ตอบแบบสอบถามพบว่า แบบสอบถามที่ใช้เวลาในการตอบระหว่าง 6-10 นาที มีความเหมาะสมมากที่สุด (ตารางที่ 1)

ตารางที่ 1 เวลาที่เหมาะสมในการตอบแบบสอบถาม (Carl McDaniel and Roger Gates, 2013 : 169)

เวลาที่เหมาะสมในการตอบแบบสอบถาม	ร้อยละของผู้เห็นด้วย
น้อยกว่า 2 นาที	2
2-5 นาที	21
6-10 นาที	44
11-15 นาที	21
16-25 นาที	3
26 นาทีขึ้นไป	0
อื่น ๆ	9

• **การใช้ข้อคำถามที่ไม่เหมาะสม** นักวิจัยมักเป็นผู้คิดข้อคำถามด้วยตนเอง บางครั้งจึงใช้คำศัพท์เฉพาะเนื่องจากมีความเคยชิน ทำให้ผู้ตอบแบบสอบถามไม่เข้าใจคำถาม ส่งผลให้เกิดข้อมูลขาดหาย เพื่อป้องกันปัญหาดังกล่าว นักวิจัยควรนำแบบสอบถามไปทดลองกับกลุ่มตัวอย่าง และสอบถามความคิดเห็นต่อข้อคำถามแต่ละข้อ เพื่อนำข้อมูลที่ได้มาปรับปรุงข้อคำถาม

• **การกำหนดกลุ่มตัวอย่างที่กว้างเกินไป** กลุ่มตัวอย่างที่มีความหลากหลายสูง ทำให้กลุ่มตัวอย่างมีความรู้และความสามารถในการให้ข้อมูลแตกต่างกัน ซึ่งอาจเป็นสาเหตุทำให้เกิดข้อมูลขาดหาย ดังนั้นนักวิจัยจึงควรกำหนดกลุ่มตัวอย่างที่มีลักษณะใกล้เคียงกันเพื่อป้องกันการเกิดปัญหาดังกล่าว

• **กระบวนการเก็บรวบรวมข้อมูลไม่เหมาะสม** โดยทั่วไปนักวิจัยมักไม่ได้เป็นผู้ดำเนินการเก็บรวบรวมข้อมูลด้วยตนเอง การให้ผู้อื่นดำเนินการเก็บรวบรวมข้อมูลแทน บางครั้งพบว่าผู้ดำเนินการเก็บรวบรวมข้อมูลต้องการให้งานเสร็จ จึงเร่งผู้ให้ข้อมูลจนทำให้เกิดข้อมูลขาดหาย เพื่อป้องกันปัญหาดังกล่าว นักวิจัยต้องมีกระบวนฝึกอบรมและควบคุมผู้เก็บรวบรวมข้อมูล

• **ความผิดพลาดในกระบวนการนำเข้าข้อมูล** การนำข้อมูลจากแบบสอบถามป้อนสู่โปรแกรมวิเคราะห์สถิติ เป็นสาเหตุหนึ่งที่ทำให้เกิดข้อมูลขาดหายอันเนื่องมาจากความผิดพลาดของผู้นำเข้าข้อมูล นักวิจัยจึงควรตรวจสอบความถูกต้องของข้อมูลก่อนนำข้อมูลไปวิเคราะห์

2. สาเหตุจากผู้ให้ข้อมูล หมายถึง การเกิดข้อมูลขาดหายเนื่องจากผู้ให้ข้อมูลให้ข้อมูลไม่ครบ ไม่สมบูรณ์ หรือไม่สามารถให้ข้อมูลอย่างต่อเนื่องด้วยเหตุผลหรือข้อจำกัดของผู้ให้ข้อมูลเอง ได้แก่

• **ผู้ให้ข้อมูลไม่เต็มใจให้ข้อมูล** เช่น ผู้ให้ข้อมูลมีเวลาจำกัด หรือถูกบังคับให้ข้อมูล เพื่อลดข้อมูลขาดหายจากผู้ให้ข้อมูลไม่เต็มใจให้ข้อมูล นักวิจัยควรอธิบายและสอบถามความสมัครใจในการให้ข้อมูลก่อนเก็บข้อมูล

• **กลุ่มตัวอย่างไม่สามารถให้ข้อมูลต่อเนื่อง** การวิจัยบางประเภทจำเป็นต้องใช้กลุ่มตัวอย่างกลุ่มเดิมในการเก็บข้อมูลอย่างต่อเนื่อง เช่น การสำรวจทัศนคติก่อนและหลังใช้สินค้า เพื่อให้ผลลัพธ์ที่ถูกต้องนักวิจัยจำเป็นต้องศึกษาโดยใช้กลุ่มตัวอย่างกลุ่มเดิม ดังนั้นในการคัดเลือกกลุ่มตัวอย่าง นักวิจัยต้องตรวจสอบกลุ่มตัวอย่างว่ามีความสามารถให้ข้อมูลอย่างต่อเนื่องได้หรือไม่

แนวทางการจัดการกับข้อมูลขาดหาย

ถึงแม้นักวิจัยจะมีกระบวนการป้องกันการเกิดข้อมูลขาดหาย แต่การเก็บรวบรวมข้อมูลในสถานการณ์จริง ย่อมเกิดข้อมูลขาดหายได้ Joseph F. Hair Jr. et. al. (2010 : 44-63) ได้ให้แนวทางการจัดการกับข้อมูลขาดหายโดยแบ่งเป็น 4 ขั้นตอน ดังนี้

ขั้นที่ 1 การระบุประเภทของข้อมูลขาดหาย ข้อมูลขาดหายมีหลายประเภท ได้แก่

• **ข้อมูลขาดหายแบบสามารถละเลยได้ (Ignorable missing data)** คือ ข้อมูลขาดหายที่เกิดจากข้อจำกัดหรือการออกแบบการวิจัย เช่น การสำรวจความพึงพอใจของการใช้บริการเครื่องเล่นในสวนสนุก นักวิจัยได้สำรวจความพึงพอใจของผู้ใช้บริการแยกเป็นเครื่องเล่นแต่ละชนิด ผู้ใช้บริการไม่ตอบข้อมูลความพึงพอใจของเครื่องเล่นบางชนิด เนื่องจากไม่ได้ใช้บริการ ข้อมูลขาดหายแบบสามารถละเลยได้อีกประเภทหนึ่ง คือ ข้อมูลที่ถูกตัดทิ้ง (Censored data) เนื่องจากสถานการณ์ที่ไม่ปกติ เช่น นักวิจัยต้องการศึกษาแนวโน้มการใช้จ่ายของผู้บริโภคต่อการซื้อสินค้าประเภทเครื่องใช้ไฟฟ้าเป็นเวลา 5 ปี แต่ในบางปีมีเหตุการณ์สำคัญที่ส่งผลให้การใช้จ่ายของผู้บริโภคมีรูปแบบที่ไม่ปกติ เช่น การเกิดเหตุการณ์น้ำท่วมใหญ่ ทำให้รูปแบบการซื้อสินค้าผิดปกติ ข้อมูลเหล่านี้เป็นข้อมูลที่ต้องถูกตัดทิ้ง ซึ่งทำให้เป็นข้อมูลขาดหายที่ไม่ต้องดำเนินการใด ๆ

• **ข้อมูลขาดหายที่ไม่สามารถละเลยได้ (Missing data processes that are not ignorable)** คือ ข้อมูลขาดหายที่เกิดจากกระบวนการที่รู้เหตุผล เช่น การลงรหัสข้อมูลผิดพลาด การไม่ควบคุมการเก็บข้อมูลให้ครบถ้วน หรือข้อมูลขาดหายเกิดจากกระบวนการที่ไม่รู้เหตุผล เช่น ผู้ให้ข้อมูลไม่ยอมให้ข้อมูล หรือให้ข้อมูลไม่ครบถ้วน ข้อมูลขาดหายประเภทนี้นักวิจัยต้องพิจารณาดำเนินการแก้ไข ไม่สามารถละเลยได้

ขั้นที่ 2 การระบุขอบเขตหรือขนาดของข้อมูลขาดหาย เป็นขั้นตอนพิจารณาขอบเขตและรูปแบบของข้อมูลขาดหาย นักวิจัยต้องวิเคราะห์ข้อมูลแบบตาราง (Tabulating) คือ ร้อยละของตัวแปรที่ขาดหายไปในแต่ละตัวอย่าง และจำนวนตัวอย่างที่ขาดหายไปในแต่ละตัวแปร หลักอย่างง่ายในการพิจารณา คือ หากข้อมูลขาดหายน้อยกว่าร้อยละ 10 สามารถละเลยได้ แต่ข้อมูลขาดหายดังกล่าวต้องเกิด

ขึ้นแบบสุ่ม เช่น ไม่ได้เกิดในข้อคำถามใดข้อคำถามหนึ่งเป็นพิเศษ นอกจากนี้จะต้องมีจำนวนข้อมูลที่สมบูรณ์มากพอสำหรับการวิเคราะห์โดยสถิติที่กำหนด บางกรณีนักวิจัยต้องการตัดตัวอย่างที่มีข้อมูลขาดหายหรือตัดตัวแปรที่มีข้อมูลขาดหาย นักวิจัยการตลาดควรพิจารณาในประเด็นสำคัญก่อนตัดข้อมูลออก คือ ข้อมูลที่ตัดออกต้องน้อยกว่าร้อยละ 15 และเมื่อตัดข้อมูลออกแล้วข้อมูลที่เหลืออยู่ต้องเพียงพอต่อการวิเคราะห์ข้อมูล

ขั้นที่ 3 การวินิจฉัยการกระจายตัวแบบสุ่มของข้อมูลขาดหาย หมายถึงการวิเคราะห์การกระจายตัวแบบสุ่มของข้อมูลขาดหาย โดยทั่วไปมี 3 รูปแบบ ได้แก่

• **การขาดหายแบบสุ่ม (Missing At Random : MAR)** หมายถึง รูปแบบข้อมูลที่ขาดหายขึ้นกับโครงสร้างของข้อมูล แต่ไม่ได้ขึ้นกับข้อมูลที่ขาดหายทั้งหมด หรือข้อมูลขาดหาย Y จะเกิดขึ้นแบบสุ่มในแต่ละตัวแปร X แต่หากพิจารณาเฉพาะข้อมูลขาดหาย Y จะพบว่าไม่เป็นแบบสุ่ม เช่น ตัวแปร Y คือน้ำหนักตัว ตัวแปร X คือ เพศ เมื่อพิจารณาข้อมูลของเพศชาย พบว่า เพศชายไม่ให้ข้อมูลน้ำหนักตัวมีการกระจายตัวแบบสุ่ม เมื่อพิจารณาข้อมูลของเพศหญิง พบว่า เพศหญิงไม่ให้ข้อมูลน้ำหนักตัวมีการกระจายตัวแบบสุ่ม แต่เพศหญิงมีอัตราการไม่ให้ข้อมูลน้ำหนักตัวสูงกว่าเพศชาย

• **การขาดหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely At Random : MCAR)** หมายถึง รูปแบบของข้อมูลที่ขาดหายเป็นแบบสุ่ม โดยข้อมูลที่ขาดหายไม่ขึ้นกับทั้งโครงสร้างของข้อมูลและข้อมูลที่ขาดหายทั้งหมด เช่น ตัวแปร Y คือน้ำหนักตัว ตัวแปร X คือ เพศ เมื่อพิจารณาข้อมูลของเพศชาย พบว่า เพศชายและเพศหญิงไม่ให้ข้อมูลน้ำหนักตัวมีการกระจายตัวแบบสุ่มในอัตราเดียวกัน ดังนั้นการดำเนินการกับข้อมูลขาดหายจึงสามารถทำได้โดยไม่ต้องคำนึงถึงปัจจัยด้านเพศ

• **การขาดหายแบบไม่สุ่ม (Missing not at Random : NMAR)** หมายถึง การขาดหายของข้อมูลขึ้นกับข้อมูลขาดหายเอง เช่น เช่น ตัวแปร Y คือน้ำหนักตัว ตัวแปร X คือ เพศ เมื่อพิจารณาข้อมูลขาดหายพบว่าผู้มีน้ำหนักตัวมากมักไม่ให้ข้อมูลด้านน้ำหนักตัว

ขั้นที่ 4 การเลือกวิธีเติมข้อมูล นอกจากนักวิจัยจะเลือกใช้ข้อมูลในแบบที่มีข้อมูลบางส่วนที่ขาดหาย หรือใช้การตัดข้อมูลหรือตัวแปรออก (ตามวิธีในขั้นที่ 1 และ 2) ในกรณีที่พบว่าการขาดหายของข้อมูลเป็นการขาดหายแบบสุ่ม หรือการขาดหาย

อย่างสมบูรณ์แบบสุ่ม นักวิจัยการตลาดสามารถพิจารณาเติมข้อมูลที่ขาดหายด้วยวิธีต่าง ๆ เช่น การใช้ค่าเฉลี่ยหรือค่าฐานนิยม (จำนวนที่มีความถี่สูงที่สุด) การใช้เทคนิคการพยากรณ์ โดยสร้างสมการเพื่อใช้พยากรณ์ค่าที่ขาดหาย

แนวทางการดำเนินการกับข้อมูลขาดหาย

เมื่อเกิดข้อมูลขาดหาย นักวิจัยมีแนวทางการดำเนินการกับข้อมูลขาดหายมีได้หลายแนวทางดังนี้

1. การตัดทิ้ง หมายถึง การตัดตัวอย่างที่ให้ข้อมูลไม่ครบถ้วนหรือไม่สมบูรณ์ออกจากข้อมูลทั้งหมด เช่น นักวิจัยเก็บข้อมูลมาได้ทั้งสิ้น 500 ตัวอย่าง พบว่ามี 30 ตัวอย่าง ที่ให้ข้อมูลไม่ครบถ้วนหรือไม่สมบูรณ์ นักวิจัยจึงตัดตัวอย่างนั้นออกจากการวิเคราะห์ข้อมูล การตัดตัวอย่างที่มีข้อมูลขาดหาย นักวิจัยต้องคำนึงถึงจำนวนตัวอย่างที่ตัดออกว่า จำนวนตัวอย่างที่คงเหลือหลังจากตัดตัวอย่างที่มีข้อมูลขาดหายมีความเพียงพอต่อการวิเคราะห์ข้อมูลหรือไม่ นอกจากนั้นนักวิจัยควรพิจารณาสัดส่วนของตัวอย่างที่ถูกตัดออกเทียบกับตัวอย่างที่ให้ข้อมูลครบถ้วน หากพบว่ามีสัดส่วนสูงมากเป็นการสะท้อนถึงการออกแบบแบบสอบถามหรือการสร้างข้อคำถามไม่เหมาะสมหรือมีปัญหาในการสุ่มตัวอย่างเพื่อเก็บรวบรวมข้อมูล

2. การเก็บข้อมูลเพิ่มเติม หมายถึง การตัดตัวอย่างที่ให้ข้อมูลไม่ครบถ้วนหรือไม่สมบูรณ์ออกจากข้อมูลทั้งหมด และดำเนินการเก็บข้อมูลเพิ่มเติมเพื่อทดแทนจำนวนตัวอย่างที่ถูกตัดออก นักวิจัยควรเลือกวิธีการเก็บข้อมูลเพิ่มเติม ในกรณีที่จำนวนตัวอย่างคงเหลือหลังจากตัดตัวอย่างที่มีข้อมูลขาดหายมีจำนวนน้อย ไม่เพียงพอต่อการวิเคราะห์ข้อมูล อย่างไรก็ดี การเก็บข้อมูลเพิ่มเติมทำให้มีค่าใช้จ่ายเพิ่ม และเสียเวลาในการเก็บรวบรวมข้อมูล

3. การเลือกตัวอย่างเพื่อการวิเคราะห์ข้อมูล หมายถึง การวิเคราะห์ข้อมูลโดยเลือกเฉพาะตัวอย่างที่มีข้อมูลที่ต้องวิเคราะห์ เช่น การศึกษาความพึงพอใจของผู้ใช้บริการโรงแรม เก็บตัวอย่างจำนวนทั้งสิ้น 400 ตัวอย่าง พบว่ามี 20 ตัวอย่าง ที่ไม่ตอบคำถามเรื่องความพึงพอใจในการใช้บริการห้องอาหาร 40 ตัวอย่าง ที่ไม่ตอบคำถามเรื่องความพึงพอใจการใช้บริการสระว่ายน้ำ เมื่อวิเคราะห์ค่าเฉลี่ยความพึงพอใจของผู้ใช้บริการห้องอาหารจะใช้จำนวนตัวอย่างเพียง 380 ตัวอย่าง และเมื่อวิเคราะห์ค่าเฉลี่ยความพึงพอใจของผู้ใช้บริการสระว่ายน้ำจะใช้จำนวนตัวอย่างเพียง 360 ตัวอย่าง

4. การแทนค่าข้อมูลขาดหาย หมายถึง การใช้เทคนิคเพื่อกำหนดค่าเพื่อทดแทนข้อมูลที่ขาดหาย ให้จำนวนตัวอย่างไม่ลดลง อย่างไรก็ดีการกำหนดค่าเพื่อทดแทนข้อมูลที่ขาดหายมีหลายเทคนิคด้วยกัน บางเทคนิคมีความถูกต้องน้อย บางเทคนิคมีความถูกต้องสูง นอกจากนั้นแต่ละเทคนิคยังมีระดับความยากง่ายแตกต่างกัน

เทคนิคการดำเนินการกับข้อมูลขาดด้วยวิธีแบบดั้งเดิม

เทคนิคการดำเนินการกับข้อมูลขาดหายด้วยวิธีแบบดั้งเดิม เป็นวิธีที่ได้รับความนิยมจากนักวิจัยในอดีต เนื่องจากมีความง่าย ไม่ซับซ้อน มีวิธีต่าง ๆ ดังนี้

Listwise deletion (LD)

Listwise deletion (LD) ได้แก่ การตัดข้อมูลที่พบว่ามีข้อมูลขาดหายออก เช่น การสำรวจความพึงพอใจของโรงแรมแห่งหนึ่ง ผู้วิจัยได้สำรวจความพึงพอใจของผู้ใช้บริการในด้านต่าง ๆ ได้แก่ ห้องพัก สระว่ายน้ำ ห้องออกกำลังกาย ห้องอาหาร และห้องบริการธุรกิจ จากการสุ่มตัวอย่างจำนวน 10 ตัวอย่าง พบว่า มีผู้ตอบคำถามความพึงพอใจต่อห้องพักจำนวน 10 ตัวอย่าง ความพึงพอใจต่อสระว่ายน้ำจำนวน 9 ตัวอย่าง ความพึงพอใจต่อห้องออกกำลังกายจำนวน 8 ตัวอย่าง ความพึงพอใจต่อห้องอาหารจำนวน 10 ตัวอย่าง และความพึงพอใจต่อห้องบริการธุรกิจจำนวน 9 ตัวอย่าง ผู้วิจัยเลือกวิธีดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion จึงเลือกวิเคราะห์ข้อมูลโดยคำนวณค่าเฉลี่ยความพึงพอใจเฉพาะผู้ให้ข้อมูลครบทุกคำถามเท่ากับ 6 คน ได้ผลดังตารางที่ 2

ข้อดีของวิธีนี้ คือ ง่าย และสามารถวิเคราะห์เชิงเปรียบเทียบระหว่างกลุ่มตัวอย่างได้ เนื่องจากกลุ่มตัวอย่างที่นำมาวิเคราะห์ตอบคำถามครบทุกข้อเหมือน ๆ กัน แต่วิธีนี้มีข้อเสีย คือ อำนาจในการวิเคราะห์จะลดลงเนื่องจากจำนวนตัวอย่างถูกตัดออก กรณีที่การกระจายตัวของข้อมูลขาดหายเป็นแบบการขาดหายแบบสุ่ม การตัดข้อมูลออกทำให้ข้อมูลมีความลำเอียงเนื่องจากข้อมูลของกลุ่มตัวอย่างแต่ละกลุ่มจะถูกตัดออกในอัตราที่ไม่เท่ากัน

ตารางที่ 2 การดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion

ตัวอย่างที่	คะแนนความพึงพอใจของการใช้บริการต่าง ๆ					ดำเนินการ
	ห้องพัก	สระว่ายน้ำ	ห้องออกกำลังกาย	ห้องอาหาร	ห้องบริการธุรกิจ	
1	2	3	2	2	5	-
2	3	3	5	3	4	-
3	5	5	1	5	ข้อมูลขาดหาย	ตัดออก
4	5	5	ข้อมูลขาดหาย	5	1	-
5	3	4	5	4	4	-
6	5	ข้อมูลขาดหาย	1	5	2	ตัดออก
7	1	2	3	2	5	-
8	2	3	5	3	4	-
9	5	4	ข้อมูลขาดหาย	5	1	ตัดออก
10	3	3	4	3	4	-
ค่าเฉลี่ย	**2.3**	**3.0**	**4.0**	**2.8**	**4.3**	**-**

Pairwise deletion (PD)

Pairwise deletion (PD) ได้แก่ การวิเคราะห์ข้อมูลโดยใช้เฉพาะข้อมูลที่มีอยู่ จากตัวอย่างการสำรวจความพึงพอใจของโรงแรม หากผู้วิจัยเลือกดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Pairwise deletion และคำนวณค่าเฉลี่ยความพึงพอใจของผู้ใช้บริการด้านต่าง ๆ ดังนี้

การคำนวณค่าเฉลี่ยความพึงพอใจด้านห้องพัก ไม่มีการตัดตัวอย่างออก เนื่องจากไม่มีข้อมูลขาดหาย พบว่าได้ค่าเฉลี่ยความพึงพอใจเท่ากับ 3.4 ซึ่งมีคะแนนความพึงพอใจเฉลี่ยสูงกว่าการดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion ที่มีคะแนนความพึงพอใจเฉลี่ยเท่ากับ 2.3 (วิธี Listwise deletion มีการ

ตัดตัวอย่างออกจำนวน 4 ตัวอย่าง)

การคำนวณค่าเฉลี่ยความพึงพอใจด้านสระว่ายน้ำ ถูกตัดตัวอย่างที่มีข้อมูลขาดหายออกจำนวน 1 ตัวอย่าง พบว่าได้ค่าเฉลี่ยความพึงพอใจเท่ากับ 3.6 ซึ่งมีคะแนนความพึงพอใจเฉลี่ยสูงกว่าการดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion ที่มีคะแนนความพึงพอใจเฉลี่ยเท่ากับ 3.0

การคำนวณค่าเฉลี่ยความพึงพอใจด้านห้องออกกำลังกาย ถูกตัดตัวอย่างที่มีข้อมูลขาดหายออกจำนวน 2 ตัวอย่าง พบว่าได้ค่าเฉลี่ยความพึงพอใจเท่ากับ 3.2 ซึ่งมีคะแนนความพึงพอใจเฉลี่ยต่ำกว่าการดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion ที่มีคะแนนความพึงพอใจเฉลี่ยเท่ากับ 4.0

การคำนวณค่าเฉลี่ยความพึงพอใจด้านห้องอาหาร ไม่มีการตัดตัวอย่างออกเนื่องจากไม่มีข้อมูลขาดหาย พบว่าได้ค่าเฉลี่ยความพึงพอใจเท่ากับ 3.7 ซึ่งมีคะแนนความพึงพอใจเฉลี่ยสูงกว่าการดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion ที่มีคะแนนความพึงพอใจเฉลี่ยเท่ากับ 2.8 (วิธี Listwise deletion มีการตัดตัวอย่างออกจำนวน 4 ตัวอย่าง)

การคำนวณค่าเฉลี่ยความพึงพอใจด้านห้องธุรกิจ ถูกตัดตัวอย่างที่มีข้อมูลขาดหายออกจำนวน 1 ตัวอย่าง พบว่าได้ค่าเฉลี่ยความพึงพอใจเท่ากับ 3.3 ซึ่งมีคะแนนความพึงพอใจเฉลี่ยต่ำกว่าการดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Listwise deletion ที่มีคะแนนความพึงพอใจเฉลี่ยเท่ากับ 4.3

จากการเปรียบเทียบการดำเนินการกับข้อมูลขาดหายด้วยวิธี Listwise deletion และวิธี Pairwise deletion พบว่า ส่งผลทำให้ได้ค่าเฉลี่ยที่แตกต่างกัน และอาจทำให้แปลผลการวิจัยผิดพลาด ดังนั้นก่อนเลือกใช้วิธีการดำเนินการกับข้อมูลขาดหลาย นักวิจัยจึงต้องพิจารณาผลกระทบกับการวิเคราะห์ข้อมูลอย่างรอบคอบ

สำหรับข้อดีของวิธี Pairwise deletion คือ การวิเคราะห์ข้อมูลจะใช้ข้อมูลที่เก็บรวบรวมได้ทั้งหมด แต่มีข้อเสีย คือ ไม่สามารถวิเคราะห์เชิงเปรียบเทียบ เนื่องจากกลุ่มตัวอย่างมีจำนวนไม่เท่ากัน และกลุ่มตัวอย่างให้ข้อมูลที่แตกต่างกัน

ตารางที่ 3 การดำเนินการกับข้อมูลขาดหายโดยใช้วิธี Pairwise deletion

ตัวอย่างที่	คะแนนความพึงพอใจของการใช้บริการต่าง ๆ					ดำเนินการ
	ห้องพัก	สระว่ายน้ำ	ห้องออกกำลังกาย	ห้องอาหาร	ห้องบริการธุรกิจ	
1	2	3	2	2	5	-
2	3	3	5	3	4	-
3	5	5	1	5	ข้อมูลขาดหาย	-
4	5	5	ข้อมูลขาดหาย	5	1	-
5	3	4	5	4	4	-
6	5	ข้อมูลขาดหาย	1	5	2	-
7	1	2	3	2	5	-
8	2	3	5	3	4	-
9	5	4	ข้อมูลขาดหาย	5	1	-
10	3	3	4	3	4	-
ค่าเฉลี่ย	**3.4**	**3.6**	**3.2**	**3.7**	**3.3**	

Single imputation

Single imputation ได้แก่ การแทนค่าข้อมูลขาดหายโดยใช้ค่าเดียว ผู้วิจัยสามารถเลือกค่าที่ใช้แทนข้อมูลขาดหายได้ดังนี้

การแทนค่าด้วยค่าเฉลี่ยเลขคณิตหรือค่าฐานนิยม (Mean/Mode substitution) ได้แก่ การคำนวณค่าเฉลี่ยเลขคณิตหรือค่าฐานนิยมจากข้อมูลทั้งหมด และนำค่าดังกล่าวมาแทนค่าของข้อมูลขาดหาย ข้อดีของวิธีนี้ คือ ทำให้จำนวนตัวอย่างไม่ลดลง แต่มีข้อเสีย คือ ข้อมูลไม่มีความหลากหลาย เนื่องจากข้อมูลขาดหายถูกแทนด้วยค่าเดียวกันทั้งหมด และไม่เหมาะสมที่จะนำไปวิเคราะห์หาความสัมพันธ์ เนื่องจากการแทนค่าข้อมูลขาดหายไม่ได้คำนึงถึงความสัมพันธ์ระหว่างตัวแปร

การแทนค่าด้วยตัวแทน (Dummy variable adjustment) ได้แก่ การสร้างตัวแปรที่ระบุข้อมูลขาดหาย เช่น หากมีข้อมูลขาดหายให้แทนค่าด้วย 1 หากมีข้อมูลครบถ้วนให้แทนค่าด้วย 0 เพื่อใช้ในการวิเคราะห์ความถดถอยเชิงพหุ (Multiple Regression Analysis) อย่างไรก็ดีการแทนค่าด้วยตัวแทน ทำให้เกิดปัญหาความลำเอียงในการหาค่าความสัมพันธ์ระหว่างตัวแปร

วิธีที่ Single imputation ที่นิยมใช้มากที่สุด คือ การแทนข้อมูลขาดหายด้วยค่าเฉลี่ย จากตัวอย่างการสำรวจความพึงพอใจของโรงแรม เมื่อแทนข้อมูลขาดหายด้วยค่าเฉลี่ย และคำนวณค่าเฉลี่ยความพึงพอใจของผู้ใช้บริการได้ผลแสดงดังตารางที่ 4

ตารางที่ 4 การดำเนินการกับข้อมูลขาดหายโดยแทนค่าข้อมูลขาดหายโดยใช้ค่าเฉลี่ย

ตัวอย่างที่	คะแนนความพึงพอใจของการใช้บริการต่าง ๆ					ดำเนินการ
	ห้องพัก	สระว่ายน้ำ	ห้องออกกำลังกาย	ห้องอาหาร	ห้องบริการธุรกิจ	
1	2	3	2	2	5	-
2	3	3	5	3	4	-
3	5	5	1	5	3.33	แทนค่า
4	5	5	3.25	5	1	แทนค่า
5	3	4	5	4	4	-
6	5	3.56	1	5	2	แทนค่า
7	1	2	3	2	5	-
8	2	3	5	3	4	-
9	5	4	3.25	5	1	แทนค่า
10	3	3	4	3	4	-
ค่าเฉลี่ย	**3.4**	**3.6**	**4.1**	**3.7**	**3.3**	

เมื่อพิจารณาจากค่าเฉลี่ยที่คำนวณได้หลังจากการดำเนินการกับข้อมูลขาดหายโดยใช้เทคนิคที่แตกต่างกัน ได้แก่ Listwise deletion, Pairwise deletion และ Single imputation จะพบว่าได้ค่าเฉลี่ยที่แตกต่างกันมาก ทั้งนี้เนื่องจากการกระจาย

ตัวของข้อมูลขาดหายไม่เป็นแบบสุ่มอย่างสมบูรณ์ ทำให้การตัดตัวอย่างบางตัวอย่างออก การตัดค่าที่ขาดหายออก หรือการแทนค่าด้วยค่าเฉลี่ย เกิดความคลาดเคลื่อนสูง

Regression Imputation

Regression Imputation ได้แก่ การใช้เทคนิคการวิเคราะห์ความถดถอยเชิงพหุเพื่อพยากรณ์ค่าของข้อมูลขาดหาย และนำค่าที่ได้จากการพยากรณ์ไปแทนค่าข้อมูลขาดหาย เป็นวิธีที่มีความซับซ้อนมากกว่าวิธีที่ผ่านมา นักวิจัยต้องพยากรณ์ค่าที่จะใช้แทนข้อมูลที่ขาดหายโดยมีสมมติฐานว่า ค่าแต่ละค่ามีความสัมพันธ์กัน นั่นหมายถึงการกระจายตัวของข้อมูลขาดหายเป็นแบบสุ่ม การใช้ค่าที่ได้จากการพยากรณ์แทนค่าข้อมูลขาดหายมีข้อดี คือ สามารถใช้ข้อมูลที่เก็บรวบรวมได้ทุกข้อมูล แต่มีข้อเสีย คือ ค่าที่ได้จากการพยากรณ์ไม่แม่นยำหากข้อมูลขาดหายมีการกระจายตัวแบบสุ่มอย่างสมบูรณ์

ตัวอย่าง การสำรวจระดับ IQ และความสามารถในการทำงาน จากการสำรวจแสดงดังตารางที่ 5 พบว่ามีข้อมูลขาดหายจำนวน 10 ตัวอย่าง ผู้วิจัยเชื่อว่า IQ และความสามารถในการทำงาน มีความสัมพันธ์กัน จึงใช้เทคนิคการวิเคราะห์ความถดถอยเชิงพหุเพื่อสร้างสมการพยากรณ์ค่าความสามารถในการทำงานเพื่อใช้แทนข้อมูลที่ขาดหาย สำหรับข้อมูล IQ 78-96 โดยใช้การสร้างสมการพยากรณ์จากข้อมูลที่สมบูรณ์สำหรับ IQ 78-96

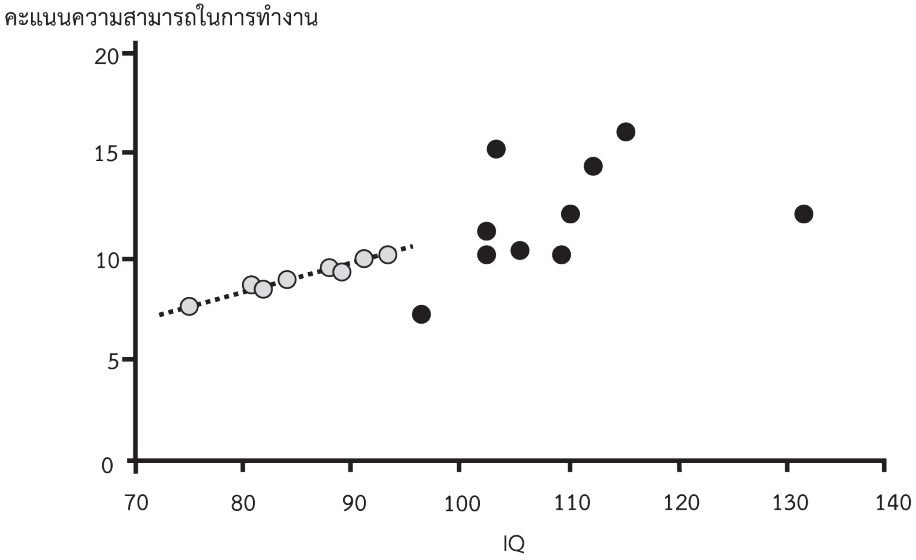
ตารางที่ 5 การเก็บข้อมูลจากพนักงาน (ปรับปรุงจาก Craig K. Enders, 2010)

ข้อมูลที่สมบูรณ์		ข้อมูลที่ขาดหาย
IQ	คะแนนความสามารถในการทำงาน	ความสามารถในการทำงาน
78	9	ข้อมูลขาดหาย
84	13	ข้อมูลขาดหาย
84	10	ข้อมูลขาดหาย
85	8	ข้อมูลขาดหาย
87	7	ข้อมูลขาดหาย
91	7	ข้อมูลขาดหาย
92	9	ข้อมูลขาดหาย
94	9	ข้อมูลขาดหาย
94	11	ข้อมูลขาดหาย
96	7	ข้อมูลขาดหาย
96	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

จากการใช้เทคนิคการวิเคราะห์ความถดถอยเชิงพหุสามารถสร้างสมการเพื่อพยากรณ์ค่าความสามารถในการทำงานดังนี้

คะแนนความสามารถในการทำงาน = -2.065 + 0.123 (IQ)

การใช้สมการพยากรณ์ข้างต้นคำนวณคะแนนความสามารถในการทำงาน แสดงดังแผนภาพที่ 1 และได้ผลลัพธ์ดังตารางที่ 6



แผนภาพที่ 1 รูปแบบการพยากรณ์ค่าโดยวิธี Regression Imputation (Craig K. Enders, 2010)

ตารางที่ 6 การเปรียบเทียบข้อมูลที่เก็บได้กับข้อมูลจากสมการพยากรณ์

IQ	คะแนนความสามารถในการทำงาน	
	คะแนนที่เก็บได้	คะแนนจากสมการพยากรณ์
78	9	7.53
84	13	8.27
84	10	8.27
85	8	8.39
87	7	8.64
91	7	9.13
92	9	9.25
94	9	9.50
94	11	9.50
96	7	9.74

(ปรับปรุงจาก Craig K. Enders, 2010)

เทคนิคการดำเนินการกับข้อมูลขาดหายโดยใช้การคำนวณที่ซับซ้อน

ในอดีตนักวิจัยนิยมเลือกการดำเนินการกับข้อมูลขาดหายโดยใช้วิธีแบบดั้งเดิม ได้แก่ Listwise deletion, Pairwise deletion, Single imputation และ Regression Imputation แต่เนื่องจากผลที่ได้มีความถูกต้องน้อย และมีความลำเอียงสูง ปัจจุบันนักวิจัยจึงนิยมเลือกใช้วิธีการดำเนินการกับข้อมูลขาดหายโดยใช้การคำนวณแบบซับซ้อน ซึ่งสามารถคำนวณได้จากโปรแกรมวิเคราะห์ทางสถิติ ทำให้ได้ผลที่มีความถูกต้องสูง และมีความลำเอียงต่ำ ตัวอย่างการดำเนินการกับข้อมูลขาดหายโดยใช้การคำนวณที่ซับซ้อนจากโปรแกรมวิเคราะห์สถิติมีดังนี้

Maximum likelihood estimation (MLE)

Maximum likelihood estimation (MLE) ได้แก่ การแทนที่ข้อมูลที่ขาดหายโดยใช้ค่าคาดหวังด้วยวิธีการประมาณค่าแบบความน่าจะเป็นสูงสุด โดยระบุกลุ่มของพารามิเตอร์ (Parameter) ที่ให้ค่า log-likelihood สูงสุด ข้อดีของวิธีนี้คือ

การใช้ข้อมูลได้ครบทุกตัวอย่าง ทั้งมีข้อมูลขาดหายและไม่มีข้อมูลขาดหาย และไม่เกิดความลำเอียงหากข้อมูลขาดหายมีการกระจายตัวทั้งแบบการขาดหายแบบสุ่มและมีการกระจายตัวแบบสุ่มอย่างสมบูรณ์

Expectation-maximization (EM) algorithm

Expectation-maximization (EM) algorithm ได้แก่ การคำนวณโดยใช้พื้นฐาน Maximum Likelihood Estimation โดยใช้วิธีประมาณค่าพารามิเตอร์ ประกอบด้วย 2 ขั้นตอน ได้แก่ ขั้นตอนประมาณค่าคาดหวัง (Expectation : E step) ได้แก่ การประมาณค่า log-likelihood ของฟังก์ชันพารามิเตอร์ และขั้นตอนการใช้ค่าสูงสุด (Maximization : M step) เป็นขั้นตอนการแทนค่าขาดหายด้วยค่าที่ได้จากขั้นตอนประมาณค่าคาดหวัง และทำการประมาณค่าคาดหวังซ้ำเพื่อเปรียบเทียบจนได้ค่าที่เปลี่ยนแปลงน้อยมาก และใช้ค่านั้นแทนข้อมูลที่ขาดหาย ข้อดีของวิธีนี้คือ การใช้ข้อมูลได้ครบทุกตัวอย่าง ทั้งมีข้อมูลขาดหายและไม่มีข้อมูลขาดหาย และไม่เกิดความลำเอียงหากข้อมูลขาดหายมีการกระจายตัวทั้งแบบการขาดหายแบบสุ่มและมีการกระจายตัวแบบสุ่มอย่างสมบูรณ์

Multiple Imputation (MI)

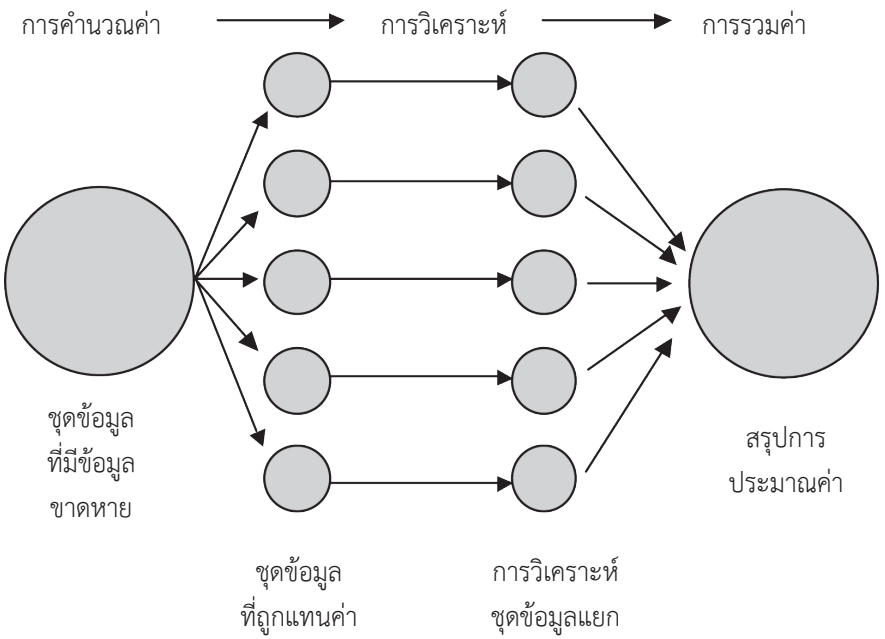
Multiple Imputation (MI) วิธีนี้มีแนวคิดใช้การแทนค่าข้อมูลขาดหายโดยยอมรับความไม่แน่นอนของค่าที่ใช้แทน ภายใต้เงื่อนไขการกระจายของข้อมูลขาดหายเป็นแบบสุ่ม วิธีนี้ประกอบด้วย 3 ขั้นตอน (แผนภาพที่ 2) ดังนี้

ขั้นตอนที่ 1 คือ ขั้นตอนการใช้เทคนิคการวิเคราะห์ความถดถอยเชิงพหุเพื่อพยากรณ์ค่า และนำค่าที่ได้จากการพยากรณ์ไปแทนข้อมูลที่ขาดหายได้เป็นชุดข้อมูล แต่ทำซ้ำเพื่อให้ได้ชุดข้อมูลหลาย ๆ ชุด โดยทั่วไปจะได้ชุดข้อมูล 5-10 ชุด

ขั้นตอนที่ 2 คือ การวิเคราะห์ข้อมูลแต่ละชุดแยกกัน เพื่อประมาณค่าพารามิเตอร์จากข้อมูลแต่ละชุด

ขั้นตอนที่ 3 คือ การรวบรวมผลที่ได้มาสรุปค่าที่จะใช้แทนข้อมูลขาดหายทั้งหมด

ข้อดีของวิธีนี้ คือ มีความถูกต้องสูงเนื่องจากใช้การคำนวณหลายครั้งเพื่อให้ได้ค่าหลาย ๆ ค่า และนำค่ามาสรุปเพื่อให้ได้ค่าที่ดีที่สุด และยังคำนึงถึงความสัมพันธ์ของความหลากหลายของกลุ่มตัวอย่าง แต่วิธีนี้มีข้อเสีย คือ มีความยุ่งยากในการคำนวณ



แผนภาพที่ 2 ขั้นตอนการการประมาณค่าโดยวิธี Multiple Imputation

Full information maximum-likelihood (FIML)

เป็นวิธีการประมาณค่าสำหรับข้อมูลขาดหายโดยใช้แบบจำลองสมการเชิงโครงสร้าง (Structural Equation Modeling) วิธีนี้ใช้การประมาณค่าพารามิเตอร์แทนการคำนวณค่าข้อมูลภายใต้สมมติฐานการกระจายตัวของข้อมูลเป็นปกติ โดยการกระจายตัวของข้อมูลขาดหายเป็นแบบสุ่มและแบบสุ่มอย่างสมบูรณ์ วิธีนี้มีข้อดี คือ เป็นการประมาณค่าจากความแตกต่างของอัตราการขาดหายของข้อมูล ขนาดของกลุ่มตัวอย่าง และรูปแบบการกระจายตัว จึงได้ผลลัพธ์ที่ถูกต้องและมีความยืดหยุ่นของข้อมูล แต่มีข้อเสีย คือ การสร้างแบบจำลองสมการเชิงโครงสร้างมีความยุ่งยาก และการพิจารณาปรับสมการเชิงโครงสร้างอาจทำให้เกิดความคลาดเคลื่อน อย่างไรก็ดีหากการปรับสมการเชิงโครงสร้างทำได้อย่างถูกต้อง จะทำให้ได้ผลลัพธ์ที่ไม่มีความลำเอียงและมีความถูกต้องสูง

แนวโน้มการเลือกใช้เทคนิคการดำเนินการกับข้อมูลขาดหาย

Yiran Dong and Chao-Ying Joanne Peng (2013) ได้สรุปแนวโน้มการเลือกใช้เทคนิคเพื่อดำเนินการกับข้อมูลขาดหายของนักวิจัยระหว่าง ค.ศ.1998-2004 พบว่า นักวิจัยเลือกใช้วิธี LD ลดลงจากร้อยละ 80.7 เป็นร้อยละ 21.7 วิธี PD ลดลงจากร้อยละ 17.3 เป็นร้อยละ 6.5 วิธี FIML เพิ่มขึ้นจากร้อยละ 0 เป็นร้อยละ 26.1 วิธี EM เพิ่มจากร้อยละ 1.0 เป็นร้อยละ 8.7 และวิธี MI เพิ่มจากร้อยละ 0 เป็นร้อยละ 6.5 จากแนวโน้มดังกล่าวจะเห็นได้ว่าในอดีตนักวิจัยเลือกการดำเนินการกับข้อมูลขาดหายโดยวิธีแบบง่าย ซึ่งทำให้ได้ผลลัพธ์ที่มีความถูกต้องน้อย ต่อมาจึงเริ่มใช้วิธีการที่ซับซ้อนและได้ผลลัพธ์ที่มีความถูกต้องสูงขึ้น

เทคนิคการดำเนินการกับข้อมูลขาดหายในโปรแกรมวิเคราะห์สถิติ

ภาควิชาสถิติและข้อมูลทางวิทยาศาสตร์ มหาวิทยาลัย Texas at Austin ได้รวบรวมข้อมูลการดำเนินการกับข้อมูลขาดหายจากโปรแกรมวิเคราะห์สถิติ พบว่า โปรแกรมวิเคราะห์สถิติมีรูปแบบการดำเนินการกับข้อมูลขาดหายที่แตกต่างกัน แสดงดังตารางที่ 7

ตารางที่ 7 การดำเนินการกับข้อมูลขาดหายในโปรแกรมวิเคราะห์สถิติ

ชื่อโปรแกรม	วิธี	สมมติฐาน	คำแนะนำ
Amelia	MI	MAR	ใช้งานง่ายถึงปานกลาง ไม่มีค่าใช้จ่าย
SAS Base	แทนค่าโดยใช้ค่าเฉลี่ย	MCAR	ใช้งานง่าย แต่แนะนำสำหรับมีข้อมูลขาดหายไม่เกินร้อยละ 5
SAS/STAT	MI	MAR	ใช้งานยาก
SAS/IML	MI	MAR	ใช้งานยาก
Paul Allison's SAS	MI	MAR	ใช้งานยาก
SAS EM	EM	MAR	ใช้งานยาก

ตารางที่ 7 การดำเนินการกับข้อมูลขาดหายในโปรแกรมวิเคราะห์สถิติ (ต่อ)

ชื่อโปรแกรม	วิธี	สมมติฐาน	คำแนะนำ
SPSS Base	แทนค่าโดยใช้ค่าเฉลี่ย	MCAR	ใช้งานง่าย แต่แนะนำสำหรับมีข้อมูลขาดหายไม่เกินร้อยละ 5
SPSS MVA add-in module	EM	MAR	ใช้งานง่าย
AMOS	MLE	MAR	ใช้งานง่าย
MX	MLE	MAR	ใช้งานง่าย ไม่มีค่าใช้จ่าย
NORM	MI	MAR	ใช้งานค่อนข้างยาก ไม่มีค่าใช้จ่าย
SOLAS	MI	MAR และ MCAR สามารถเลือกได้	ใช้งานค่อนข้างง่าย

(ปรับปรุงจาก https://stat.utexas.edu/software-faqs/general)

บทสรุป

การเกิดข้อมูลขาดหายเป็นสิ่งที่พบได้เป็นปกติในการเก็บรวบรวมข้อมูล ซึ่งอาจเกิดจากกระบวนการวิจัย เช่น การกำหนดประเด็นการวิจัยที่กว้างเกินไป การออกแบบแบบสอบถามที่ใช้เวลาตอบมากเกินไป การใช้ข้อคำถามที่ไม่เหมาะสม การกำหนดกลุ่มตัวอย่างที่กว้างเกินไป กระบวนการเก็บรวบรวมข้อมูลไม่เหมาะสม หรือจากความผิดพลาดในกระบวนการนำเข้าข้อมูล ในบางกรณีข้อมูลขาดหายเกิดจากผู้ให้ข้อมูลเอง เช่น ผู้ให้ข้อมูลไม่เต็มใจให้ข้อมูล หรือกลุ่มตัวอย่างไม่สามารถให้ข้อมูลต่อเนื่อง ถึงแม้ว่าการเกิดข้อมูลขาดหายจะเป็นเรื่องที่เกิดได้เป็นปกติ แต่หากนักวิจัยมีความรอบคอบในการออกแบบการวิจัย สามารถช่วยลดจำนวนข้อมูลขาดหายได้

ในอดีตเมื่อเกิดข้อมูลขาดหายในงานวิจัย นักวิจัยบางกลุ่มไม่ดำเนินการกับข้อมูลขาดหาย แต่นักวิจัยบางกลุ่มเลือกใช้วิธีดำเนินการกับข้อมูลขาดหายแบบง่ายโดยใช้วิธีตัดข้อมูลออก เช่น วิธี Listwise deletion และ Pairwise deletion ซึ่งวิธีดังกล่าว

มีความถูกต้องน้อย ภายหลังนักวิจัยนิยมใช้วิธีดำเนินการกับข้อมูลขาดหายที่มีความถูกต้องมากขึ้น เช่น Full information maximum-likelihood, Expectation-maximization algorithm และ Multiple Imputation

เมื่อเกิดข้อมูลขาดหายในการวิจัย นักวิจัยควรดำเนินการกับข้อมูลขาดหาย ไม่ควรปล่อยให้มีข้อมูลขาดหาย หรือไม่ควรตัดข้อมูลขาดหายทิ้งไป วิธีการแทนค่าข้อมูลขาดหายมีหลายวิธีด้วยกัน วิธีแทนค่าด้วยค่าเฉลี่ยเลขคณิตเป็นวิธีที่ง่ายแต่มีความถูกต้องต่ำ ดังนั้นผู้วิจัยควรดำเนินการกับข้อมูลขาดหายด้วยวิธีที่มีความถูกต้องสูง เช่น Full information maximum-likelihood, Expectation-maximization algorithm และ Multiple Imputation โดยนักวิจัยสามารถเลือกใช้โปรแกรมวิเคราะห์ที่เหมาะสม เพื่อให้ได้ผลการวิจัยที่ถูกต้อง

เอกสารอ้างอิง

Angela M. Wood, Ian R. White and Simon G. Thompson. “Are Missing Outcome Data Adequately Handled ?.” **A Review of Published Randomized Controlled Trials in Major Medical Journals Clinical Trial** 1 (2004) : 368-376.

Carl McDaniel and Roger Gates. 2013. **Marketing Research.** Ninth edition. New York : Wiley.

Chao-Ying Joanne Peng et al. “Advances in missing data methods and implications for educational research.” In : Sawilowsky SS, editor. **Real data analysis**. Charlotte, North Carolina : Information Age Pub ; 2006. pp. 31–78.

Craig K. Enders. 2010. **Applied Missing Data Analysis**. New York : Guilford Press.

Hair Joseph .F, et al. 2010. **Multivariate Data Analysis.** Seventh edition. New York : Prentice Hall.

Yiran Dong and Chao-Ying Joanne Peng. Principled missing data methods for researchers. http://www.ncbi.nlm.nih.gov/. 8 May 2015.