บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทที่ 2 นี้ แบ่งเป็นสองส่วนคือ ส่วนที่หนึ่งกล่าวถึงความรู้พื้นฐานที่เกี่ยวข้องกับข้อมูลสูญหายได้แก่ ขนาดตัวอย่าง รูปแบบของข้อมูลสูญหายที่ถูกจำลองขึ้นในการทดลอง เทคนิคที่ใช้ในการประมาณค่าข้อมูลสูญหาย การแทนค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ การประมาณค่าพารามิเตอร์ การจำลองสถานการณ์ ส่วนที่สองเป็นการกล่าวถึงงานวิจัยที่เกี่ยวข้องต่าง ๆ ที่ได้พัฒนามาเพื่อแก้ปัญหาในเรื่องของการประมาณค่าข้อมูลสูญหาย

2.1 ความรู้พื้นฐาน

2.1.1. ประชากร (Population)

ประชากร หมายถึง กลุ่มของหน่วยตัวอย่างที่ผู้วิจัยสนใจศึกษา ประชากรอาจเป็นบุคคล องค์กร สัตว์ สิ่งของที่นำมาเป็นหน่วยศึกษาหรือปัญหาในการวิจัย ถ้าประชากรนั้นประกอบด้วยหน่วยตัวอย่างเป็นจำนวนหน่วยที่นับได้แน่นอน หรือสามารถระบุจำนวนที่แน่นอนได้ เรียกว่าประชากรจำกัด (finite population) แต่ถ้าประชากรประกอบด้วยหน่วยตัวอย่างจำนวนหน่วยที่นับไม่ได้แน่นอนหรือไม่สามารถระบุจำนวนที่แน่นอนได้เรียกว่าประชากรอนันต์ (infinite population) สำหรับงานวิจัยนี้ศึกษาเฉพาะในส่วนของประชากรจำกัดเท่านั้น

2.1.2 ตัวอย่าง (Sample)

ตัวอย่าง หมายถึง ส่วนหนึ่งของประชากรที่ผู้วิจัยสนใจศึกษา ตัวอย่างที่ดีหมายถึง ตัวอย่างที่มีลักษณะต่าง ๆ ที่สำคัญครบถ้วนเหมือนกับประชากร

ขนาดตัวอย่าง (Sample size)

ขนาดตัวอย่าง คือ จำนวนหน่วยตัวอย่างในการวิจัย เป็นสิ่งสำคัญที่ผู้วิจัยต้องกำหนดให้เหมาะสมและมีความเป็นตัวแทนที่ดีของประชากรที่ทำการศึกษาเพื่อจะช่วยให้ผลการวิจัยมีความน่าเชื่อถือ

การกำหนดขนาดตัวอย่าง

การกำหนดขนาดตัวอย่างควรมีขนาดเท่าใดนั้นผู้วิจัยควรคำนึงถึงสิ่งต่าง ๆ หลายอย่างมาประกอบกัน (Librero, 1985 อ้างถึงใน ธีรวุฒิ เอกะกุล, 2543) ดังนี้

(1) ค่าใช้จ่าย เวลา แรงงานและเครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลตัวอย่างนั้นว่ามี เพียงพอหรือไม่และคุ้มค่าเพียงใด

(2) ขนาดประชากร ถ้าประชากรมีขนาดใหญ่ มีความจำเป็นต้องเลือกตัวอย่าง ถ้าประชากรมีขนาดเล็กและสามารถที่จะศึกษาได้ควรจะศึกษาจากประชากรทั้งหมด

(3) ความเหมือนกัน ถ้าประชากรมีความเหมือนกันมากความแตกต่างของหน่วยตัวอย่างมีน้อยนั่นคือความแปรปรวนของตัวอย่างมีค่าน้อยใช้ตัวอย่างขนาดเล็กได้ ถ้าประชากรมีลักษณะไม่เหมือนกันความแตกต่างของหน่วยตัวอย่างมีค่ามากทำให้ความแปรปรวนในกลุ่มมากจำเป็นต้องใช้ตัวอย่างขนาดใหญ่เพื่อให้ครอบคลุมคุณลักษณะต่าง ๆ ของประชากร

(4) ความแม่นยำ ถ้าต้องการความแม่นยำในเรื่องที่จะศึกษาค้นคว้า ต้องใช้ตัวอย่างขนาดใหญ่ คือ ยิ่งขนาดตัวอย่างใหญ่มากเท่าใดผลการศึกษายิ่งมีความแม่นยำมากขึ้นเท่านั้น

(5) ความคลาดเคลื่อนจากการชักตัวอย่าง ความคลาดเคลื่อนที่ยอมให้เกิดขึ้นได้จากการชักตัวอย่างโดยทั่วไปแล้วมักยอมให้เกิดความคลาดเคลื่อนได้ 1% หรือ 5% (สัดส่วน 0.01 หรือ 0.05) และยังขึ้นอยู่กับความสำคัญของเรื่องที่ต้องการศึกษาด้วยถ้าปัญหามีความสำคัญมากก็ควรให้เกิดความคลาดเคลื่อนน้อยที่สุดเช่น 1% แต่ถ้ามีความสำคัญน้อยก็อาจยอมให้เกิดความคลาดเคลื่อนได้บ้าง เช่น 5% เป็นต้น

(6) ความเชื่อมั่น ผู้วิจัยต้องกำหนดความเชื่อมั่นว่าตัวอย่างที่สุ่มมานั้นมีโอกาสได้ค่าอ้างอิงไม่แตกต่างจากค่าที่แท้จริงของประชากรประมาณเท่าไร เช่น ถ้ากำหนดระดับเชื่อมั่น 95% หมายถึง ค่าอ้างอิงมีโอกาสถูกต้อง 95% มีโอกาสผิดพลาดจากค่าที่แท้จริง 5% นั่นคือ ค่าที่ได้จากกลุ่มตัวอย่าง 95 กลุ่มจาก100 กลุ่มที่สุ่มมาจากประชากรเดียวกันจะไม่แตกต่างจากค่าที่แท้จริงของประชากรซึ่งระดับความเชื่อมั่นอาจจะเพิ่มขึ้นเป็น 99% หรือลดลงเหลือ 90%

2.1.3 การประมาณค่าที่เก็บไม่ได้ (Imputation)

การประมาณค่าที่เก็บไม่ได้ คือการประมาณค่าข้อมูลสูญหาย และใช้ค่าที่ประมาณนั้นมาวิเคราะห์เชิงสถิติ Laaksonen (2000) ได้จำแนกวิธีการในการประมาณค่าเป็น 2 กลุ่ม คือ

model-donor imputation คือ การประมาณค่าที่ได้มาจากตัวแบบ เช่น ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย ตัวประมาณข้อมูลสูญหายด้วยค่าถดถอยและตัวประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน

real-donor imputation คือ การประมาณค่าที่ได้มาจากเซตของค่าสังเกต เช่น ตัวประมาณข้อมูลสูญหายด้วยค่าโคลเดค ตัวประมาณข้อมูลสูญหายด้วยค่าฮอทเดค และตัวประมาณข้อมูลสูญหายด้วยค่าใกล้สุด

2.1.4 ข้อมูลสูญหาย (Missing data)

ข้อมูลสูญหาย คือ ค่าสังเกตที่ต้องการทราบค่าแต่ไม่สามารถทราบค่าได้หรือไม่สามารถเก็บรวบรวมได้ ข้อมูลสูญหายไม่เกิดขึ้นหากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีหรือมีความเหมาะสม ข้อมูลสูญหายส่งผลต่อการวิจัยทั้งในส่วนของการวิเคราะห์และการสรุปผลตีความ โดยระดับความรุนแรงของผลกระทบนี้ขึ้นอยู่กับองค์ประกอบหลายส่วน แต่ที่สำคัญคือขนาดของข้อมูลสูญหาย ประเภทของข้อมูลสูญหายที่เกิดขึ้น และวิธีการจัดการกับข้อมูลสูญหาย (ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี, 2549)

2.1.4.1 ขนาดของข้อมูลสูญหาย (Size of missing data)

กรณีที่ไม่มีคำตอบหรือเกิดการปล่อยว่างไว้สำหรับบางรายการ นักวิจัยจำเป็นต้องตัดสินใจดำเนินการระหว่างการพิจารณาไม่เลือกใช้ข้อมูลในรายที่เกิดปัญหานี้หรือควรแทนค่าสูญหาย (imputed value) ปัจจุบันยังไม่มีเกณฑ์อย่างง่ายสำหรับใช้ในการตัดสินใจ เลือกใช้วิธีการใดวิธีการหนึ่งแต่มีแนวทางหนึ่งที่ถูกนำมาใช้กันอย่างแพร่หลายคือ การพิจารณาจากขนาดของข้อมูลสูญหาย เมื่อจำนวนของหน่วยตัวอย่างที่มีค่าของข้อมูลสูญหายมีจำนวนน้อย (อาจใช้เกณฑ์ < 5% จากหน่วยตัวอย่างที่มีขนาดใหญ่) อาจเลือกใช้วิธีการอย่างง่ายคือใช้วิธีการไม่เลือกข้อมูลที่มีปัญหามาใช้ในการวิเคราะห์ ดังนั้นในการศึกษาจึงควรพยายามลดขนาดของข้อมูลสูญหายให้เหลือจำนวนน้อยที่สุดหรือไม่เหลือเลย การหาทางป้องกันไม่ให้เกิดข้อมูลสูญหาย แนวทางที่จะลดขนาดของข้อมูลสูญหายจำเป็นต้องพิจารณาถึงสาเหตุต่าง ๆ ที่เป็นไปได้ในการเกิดข้อมูลสูญหาย

2.1.4.2 ประเภทของข้อมูลสูญหาย (Missing type of data)

ประสิทธิภาพของขั้นตอนวิธีหรือเทคนิคที่นำมาใช้ในการประมาณค่าข้อมูลสูญหายนั้นจะดีหรือไม่ส่วนหนึ่งขึ้นอยู่กับประเภทการสูญหายของข้อมูล ถ้าสามารถทราบสาเหตุที่ทำให้เกิดข้อมูลสูญหายอาจจะสามารถเติมเต็มหรือเดาข้อมูลส่วนนั้นได้ไม่ยากนัก ซึ่งในการทำงานจริงไม่อาจจะทราบได้ว่าข้อมูลสูญหายนั้นมีสาเหตุมาจากอะไรและสูญหายในลักษณะใด ดังนั้นในการทดลองต่าง ๆ จึงใช้วิธีจำลองรูปแบบของข้อมูลสูญหายในหลายลักษณะโดยจำแนกออกได้เป็น 3 ประเภท ดังต่อไปนี้

(1) การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing completely at random หรือ *MCAR*) เกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของข้อมูลที่ตำแหน่งใด ๆ นั้นไม่มีความสัมพันธ์กับค่าของข้อมูลตัวอื่น ๆ ไม่ว่าจะเป็นข้อมูลที่ทราบค่าหรือข้อมูลที่เกิดการสูญหายด้วยกันก็ตาม หมายถึงความน่าจะเป็นที่จะเกิดการสูญหายของค่าข้อมูลในทุก ๆ ตำแหน่งมีค่าเท่ากัน ในการสุ่มรูปแบบนี้พบว่าวิธีการที่ใช้ในการประมาณค่าข้อมูลนั้นสามารถนำมาประยุกต์ใช้ได้ในทุก ๆ วิธี โดยที่จะไม่ทำให้ข้อมูลเกิดการเบี่ยงเบน

สำหรับสาเหตุที่ทำให้ข้อมูลเกิดการสูญหายแบบสุ่มอย่างสมบูรณ์ อาจเกิดขึ้นเนื่องจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้อง

สำหรับข้อมูลสูญหายประเภทนี้ จัดเป็นข้อมูลสูญหายที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นจึงสามารถเลือกวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้

(2) การสูญหายแบบสุ่ม (Missing at random หรือ *MAR*) การสูญหายของค่าข้อมูลในกรณีนี้ขึ้นอยู่กับค่าของข้อมูลตัวอื่น ๆ ที่ทราบค่าข้อมูลด้วย แต่ไม่ขึ้นอยู่กับค่าข้อมูลที่สูญหายของตัวเองยกตัวอย่าง เช่น หากพบว่าเฉพาะกลุ่มผู้ได้รับการศึกษาน้อยที่ไม่ให้ความร่วมมือในการตอบคำถามเกี่ยวกับทัศนคติในการเสพยาเสพติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพติดมีค่าการสูญหายแบบสุ่มทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะบางส่วนของตัวแปรระดับการศึกษา

(3) การสูญหายแบบไม่สุ่ม (Not missing at random หรือ *NMAR*) การสูญหายของค่าข้อมูลในกรณีนี้เกิดขึ้นเมื่อความน่าจะเป็นของแถวที่มีการสูญหายที่ตำแหน่งใด ๆ ขึ้นอยู่กับค่าของข้อมูลตำแหน่งนั้น ๆ หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับค่าตัวแปรใด ๆ ในฐานข้อมูลเลยแต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้นแต่เนื่องจากตัวแปรน้ำหนักตัวตอนเริ่มต้นไม่ได้ถูกเก็บรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูล ลักษณะข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่สามารถส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูล

ในทางปฏิบัติลักษณะของข้อมูลสูญหายประเภทการสูญหายแบบสุ่มอย่างสมบูรณ์มักไม่ค่อยพบบ่อยนักส่วนใหญ่เป็นข้อมูลสูญหายประเภทการสูญหายแบบสุ่ม ดังนั้นวิธีการเชิงสถิติต่าง ๆ ที่พัฒนาขึ้นมาเพื่อแก้ปัญหาข้อมูลสูญหายมักดำเนินการภายใต้ข้อสมมุติของการสูญหายแบบสุ่ม

2.1.4.3 วิธีแก้ไขข้อมูลสูญหาย (Methods for treating missing data) จากการศึกษาพบว่าวิธีแก้ไขข้อมูลสูญหายมีหลายวิธี การเลือกใช้วิธีการใดขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้นหากเลือกวิธีการที่ไม่เหมาะสมมาใช้อาจทำให้ผลลัพธ์ผิดพลาดได้ วิธีการจัดการกับข้อมูลสูญหายที่มีการนำมาใช้มีดังนี้

(1) การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (Ignoring and discarding data) เป็นวิธีการที่แบ่งเป็น 2 กลุ่มหลัก ๆ คือ การตัดแถวที่สูญหายทิ้งไปทั้งหมด (Complete case analysis) ซึ่ง

นิยมใช้ในงานด้านสถิติ อีกวิธี คือ การตัดเฉพาะแถวที่มีการสูญหายของค่าข้อมูลทิ้ง ซึ่งก่อนจะตัดตำแหน่งใด ๆ ออกไป ต้องวิเคราะห์หาความสัมพันธ์ระหว่างข้อมูลก่อนเสมอ(Afifi & Clark,1996 อ้างถึงใน นงลักษณ์ วิรัชชัย, 2542)

(2) การตัดข้อมูลแบบลิสทไวส (Listwise data deletion) เป็นวิธีการจัดการข้อมูลสูญหายที่ง่ายมากนั่นคือไม่สนใจข้อมูลสูญหายที่เกิดขึ้นโดยวิเคราะห์ข้อมูลจากข้อมูลเฉพาะส่วนที่สมบูรณ์ แนวทางนี้มีความเหมาะสมในกรณีที่ข้อมูลสูญหายมีจำนวนน้อยมากหรือผลจากการวิเคราะห์ข้อมูลมีความชัดเจนมากซึ่งวิธีการนี้มักถูกกำหนดให้ใช้เป็นตัวเลือกอัตโนมัติ (by default) สำหรับจัดการกับข้อมูลสูญหายในโปรแกรมคอมพิวเตอร์ทางสถิติทั่ว ๆ ไปหากไม่เจาะจงเลือกใช้วิธีการอื่นใดในการจัดการกับข้อมูลสูญหาย

ข้อเสียของวิธีการตัดข้อมูลแบบลิสทไวส

(2.1) ไม่เหมาะสมที่จะนำไปใช้กับข้อมูลจริงเนื่องจากแถวของข้อมูลที่ไม่สมบูรณ์ถูกตัดทิ้งไปทำให้มีจำนวนของข้อมูลที่นำไปวิเคราะห์นั้นลดน้อยลงจนบางครั้งไม่เพียงพอต่อการนำไปวิเคราะห์

(2.2) ในกรณีที่ต้องการคำนวณหาค่าสถิติบางค่าของข้อมูลนั้น ข้อมูลต้องมีการสูญหายแบบสุ่มอย่างสมบูรณ์ถ้าข้อมูลมีการสูญหายแบบสุ่มพบว่าค่ากลางที่คำนวณมาได้อาจเกิดการเบี่ยงเบนจากความเป็นจริงได้

(3) การตัดข้อมูลแบบแพรไวส (Pairwise data deletion) เป็นวิธีการจัดการกับข้อมูลสูญหายสำหรับในกรณีที่วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ โดยวิเคราะห์ข้อมูลเฉพาะข้อมูลที่สมบูรณ์ของทั้งสองตัวแปร วิธีการนี้มีข้อดีในส่วนของการใช้งานข้อมูลได้เต็มที่และมีประสิทธิภาพ แต่มีขั้นตอนซับซ้อนกว่าวิธีการตัดข้อมูลแบบลิสทไวสเล็กน้อยและเสียเวลามากกว่าจึงได้รับความนิยมน้อยกว่า

(4) วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยเป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าในแต่ละกลุ่มย่อยของตัวแปรอื่น ซึ่งเป็นวิธีที่พัฒนามาจากการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าทั้งนี้เนื่องจากข้อสมมุติว่าค่าของข้อมูลสูญหายควรขึ้นอยู่กับลักษณะของหน่วยตัวอย่างโดยลักษณะของหน่วยตัวอย่างที่ใกล้เคียงกันควรมีค่าข้อมูลที่คล้ายคลึงกันวิธีนี้เสนอโดย Wilks ในปี ค.ศ. 1932 โดยกำหนดให้

$$y_{\cdot i} = \begin{cases} y_i & if\ i \in R \\ \overline{y}_{MI} & if\ i \in R^c \end{cases} \qquad (2.1)$$

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.1) อยู่ในรูป

$$\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_{\cdot i} = \bar{y}_{MI} \tag{2.2}$$

เมื่อ $\bar{y}_{MI} = \frac{1}{r}\sum_{i\in R} y_i$ โดยที่

y_i คือ ค่าของตัวแปรที่สนใจ y ของหน่วยสังเกตที่ i ซึ่งมีทั้งหมด r หน่วย

$y_{\cdot i}$ คือ ค่าของตัวแปรที่สนใจ y ของข้อมูลที่สมบูรณ์ r หน่วย และค่าของตัวแปรที่สนใจ y ของหน่วยสังเกตที่ i หลังทำการประมาณค่าข้อมูลสูญหาย $n-r$ หน่วย

$\bar{y}_{MI}$ คือ ค่าเฉลี่ยของตัวแปรที่สนใจ y จากข้อมูลที่สมบูรณ์ทั้งหมด r หน่วย

เป็นตัวประมาณค่าที่ไม่เอนเอียง (Unbiased) และมีค่าความแปรปรวน

$$V(\bar{y}_{MI}) = \left(\frac{1}{r} - \frac{1}{N}\right) S_Y^2 \tag{2.3}$$

(5) วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย เป็นวิธีการประมาณค่าข้อมูลสูญหายโดยการสร้างสมการถดถอยระหว่างตัวแปรที่สนใจ y กับตัวแปรช่วย x โดยใช้ข้อมูลที่สมบูรณ์ทั้งหมด r หน่วยแล้วประมาณค่าข้อมูลสูญหายด้วยสมการถดถอย สุนันทา วีรกุลเทวัญ (2544) กล่าวว่า ถึงแม้ว่าวิธีการ แทนค่าด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย ทำให้ความแปรปรวนของข้อมูลมีค่าน้อยกว่าความแปรปรวนที่ได้จากวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยแต่มีข้อเสียคือทำให้ค่าที่เกี่ยวข้องกับความสัมพันธ์ (Association) ระหว่างตัวแปรบิดเบือน (Distort) เพราะสมการถดถอยที่ใช้ประมาณค่าข้อมูลสูญหายสร้างมาจากความสัมพันธ์ระหว่างตัวแปร

(6) วิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด เป็นวิธีการแทนข้อมูลสูญหายโดยใช้ข้อมูลจากการเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกันมากที่สุดกับหน่วยตัวอย่างที่เกิดค่าสูญหาย จากนั้นแทนค่าข้อมูลสูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงกัน

(7) วิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค เป็นวิธีการแทนข้อมูลสูญหายโดยใช้ข้อมูลจากผู้ให้ข้อมูล (Donor) ในงานวิจัยหรืองานสำรวจเรื่องเดียวกันที่เพิ่งสำรวจเสร็จสิ้นใหม่ ๆ (เรียกว่า hot) ไปเป็นข้อมูลทดแทนให้แก่รายการที่สูญหาย (Roth & Switzer, 1995)

(8) วิธีการประมาณข้อมูลสูญหายด้วยค่าโคลเดค เป็นวิธีการแทนข้อมูลสูญหายโดยใช้ข้อมูลจากผู้ให้ข้อมูล ในงานวิจัยหรืองานสำรวจจากเฟ้มอื่นแต่เป็นงานสำรวจเรื่องเดียวกันในการสำรวจครั้งก่อน (เรียกว่า cold) ไปเป็นข้อมูลทดแทนให้แก่รายการที่สูญหาย

(9) วิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ เป็นวิธีการแทนข้อมูลสูญหายด้วยค่าประมาณที่นำเสนอโดย Singh and Horn (2000) โดยกำหนดให้

$$y_{\cdot i} = \begin{cases} \alpha \frac{n}{r} y_i + (1-\alpha)\hat{b}x_i & if\ i \in R \\ (1-\alpha)\hat{b}x_i & if\ i \in R^c \end{cases} \tag{2.4}$$

เมื่อ $\hat{b} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_i}$ และ α คือ ค่าที่เหมาะสมที่ถูกเลือก

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.4) อยู่ในรูป

$$\bar{y}_{COMP} = \alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{2.5}$$

เมื่อ

$\bar{x}_r$ คือ ค่าเฉลี่ยของตัวแปรช่วย x ที่มีค่าข้อมูลของตัวแปรที่สนใจ y สมบูรณ์ทั้งหมด r หน่วย

$\bar{x}_n$ คือ ค่าเฉลี่ยของตัวแปรช่วย x จากข้อมูลที่สมบูรณ์ทั้งหมด n หน่วย

$\bar{y}_r$ คือ ค่าเฉลี่ยของตัวแปรที่สนใจ y จากข้อมูลที่สมบูรณ์ทั้งหมด r หน่วย

เป็นตัวประมาณค่าที่เอนเอียง โดยมีความเอนเอียง (Biasedness) ดังนี้

$$b(\bar{y}_{COMP}) = (1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}(C_X^2 - \rho_{YX}C_Y C_X) \tag{2.6}$$

และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$MSE(\bar{y}_{COMP}) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left[(1-\alpha)^2 C_X^2 - 2(1-\alpha)\rho_{YX}C_Y C_X\right] \tag{2.7}$$

และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่เหมาะสม

$$MSE(\bar{y}_{COMP})_{opt} = M(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - \rho_{YX}\frac{C_Y}{C_X}\right)^2 \bar{Y}^2 C_X^2 \tag{2.8}$$

2.1.5 การประมาณค่าพารามิเตอร์

การประมาณค่าพารามิเตอร์ มี 2 ชนิด คือ

การประมาณค่าแบบค่าเดียว (Point estimation) คือ การประมาณค่าพารามิเตอร์ของประชากรโดยใช้ค่าสถิติเพียงค่าเดียว คำนวณได้จากตัวอย่างโดยทั่วไป การประมาณค่าพารามิเตอร์ต่าง ๆ จะทำได้โดยการเลือกตัวอย่างมาจำนวน n หน่วยเก็บรวบรวมข้อมูลจากหน่วยตัวอย่างที่เลือกมาได้จากนั้นคำนวณหาค่าของตัวสถิติแล้วนำไปใช้สำหรับการประมาณค่าของประชากร

การประมาณค่าแบบเป็นช่วง (Interval estimation) คือ การประมาณค่าของพารามิเตอร์ของประชากรออกมาเป็นช่วงซึ่งผลที่ได้มาจากการประมาณแบบค่าเดียว ช่วงของการประมาณค่า พารามิเตอร์ดังกล่าวเรียกว่าช่วงความเชื่อมั่น (confidence interval) สำหรับงานวิจัยนี้ศึกษาเฉพาะการประมาณค่าแบบค่าเดียวเท่านั้น

2.1.6 การจำลองสถานการณ์ (Simulation) เป็นการรวบรวมวิธีการต่าง ๆ ที่ใช้ในการจำลองสถานการณ์จริงหรือพฤติกรรมของระบบต่าง ๆ มาไว้บนคอมพิวเตอร์ เพื่อศึกษาการไหลของกิจกรรมในรูปแบบต่าง โดยมีการเก็บข้อมูลและวิเคราะห์หารูปแบบที่ถูกต้องจากโปรแกรมคอมพิวเตอร์เพื่อปรับปรุงในอนาคต (Kelton, Sadowski, and Sturrock, 2003)

เนื่องจากในการปฏิบัติงานจริงไม่สามารถทำการทดลองหรือปรับเปลี่ยนกระบวนการทำงานได้จนกว่าจะมองเห็นถึงประโยชน์ที่ได้รับ ดังนั้นการจำลองสถานการณ์จะช่วยให้สามารถวิเคราะห์สภาพที่เป็นอยู่ในปัจจุบันของระบบ ช่วยหาแนวทางหรือทางเลือก (alternative) ที่เหมาะสมก่อนนำไปใช้กับสถานการณ์หรือการปฏิบัติงานจริง ซึ่งจะช่วยให้ลดความเสี่ยงในการเกิดความผิดพลาดหรือความล้มเหลวได้ นอกจากนี้ยังช่วยให้ประหยัดค่าใช้จ่ายและเวลาอีกด้วย (Maria, 1997)

คุณลักษณะสำคัญของแบบจำลองสถานการณ์

แบบจำลองสถานการณ์มีคุณลักษณะที่แตกต่างไปจากแบบจำลองชนิดอื่น ๆ ทั่วไปดังนี้

(1) แบบจำลองสถานการณ์ทุกชนิดต้องมีการตรวจสอบความถูกต้องก่อนเป็นอันดับแรกเพื่อไม่ให้เกิดข้อผิดพลาด มีการตรวจสอบทั้งตรรกะ (Logic) และการคำนวณว่าถูกต้องหรือไม่

(2) มีการตรวจสอบว่าผลที่ได้ต้องอยู่ในขอบเขตของผลลัพธ์ที่คาดคะเนไว้ และแบบจำลองนั้นทำงานอย่างถูกต้องโดยสามารถนำผลลัพธ์นั้นมาวิเคราะห์ได้

(3) ลดความเบี่ยงเบนโดยใช้ค่าสุ่มเดียวกันเพื่อลดความแปรผัน และเพิ่มความถูกต้องเมื่อเปรียบเทียบกับองค์ประกอบที่ต่างกันได้

(4) มีลักษณะเป็นการเลียนแบบสถานการณ์จริงมากกว่าเป็นการนำเสนอสถานการณ์จริง

(5) มีลักษณะเป็นการคาดการณ์สถานการณ์จริงที่จะเกิดขึ้นภายใต้เงื่อนไขต่าง ๆ กัน

(6) เป็นแบบจำลองที่ใช้กับปัญหาที่มีความซับซ้อนสูง

ประโยชน์ของการจำลองสถานการณ์

การที่ระบบสนับสนุนการตัดสินใจมีการใช้แบบจำลองสถานการณ์เพื่อเลียนแบบสถานการณ์ปัญหาต่าง ๆ ก่อให้เกิดประโยชน์ ดังต่อไปนี้

(1) แบบจำลองสถานการณ์เป็นทฤษฎีที่มีการใช้งานเพื่อคาดการณ์เหตุการณ์ในอนาคตอย่างตรงไปตรงมา

(2) แบบจำลองสถานการณ์สามารถทำงานที่มีเวลาเข้าไปเกี่ยวข้องเป็นจำนวนมาก ๆ ได้ดี

(3) แบบจำลองสถานการณ์ค่อนข้างเป็นการอธิบายให้เห็นเป็นรูปร่างมากกว่าการใช้เป็นเครื่องมือธรรมดา

(4) ผู้สร้างระบบการตัดสินใจสามารถติดต่อกับผู้ใช้ได้เพื่อรับรู้เรื่องราวเกี่ยวกับปัญหาได้อย่างลึกซึ้ง

(5) สามารถสร้างแบบจำลองสถานการณ์ที่มาจากมุมมองของผู้ใช้ได้

(6) แบบจำลองถูกสร้างขึ้นเฉพาะเหตุการณ์ที่มาจากมุมมองของผู้ใช้ได้

(7) แบบจำลองสถานการณ์สามารถจัดการกับปัญหาได้มากมายหลากหลายชนิด เช่น การจัดการกับคลังสินค้า (Inventory) และการจัดการทรัพยากรบุคคล (Human Resource) อีกทั้งสามารถทำหน้าที่ในเชิงบริหารระดับสูงอีกด้วย เช่น การวางแผนการต่าง ๆ ในระยะยาว ดังนั้นผู้ใช้สามารถ นำแบบจำลองสถานการณ์มาใช้จัดการสิ่งต่าง ๆ ได้ทุกช่วงเวลา

(8) สามารถทดลองป้อนตัวแปรที่แตกต่างกันตามแต่ละเหตุการณ์ลงในแบบจำลองเพื่อดูผลลัพธ์ที่เป็นทางเลือกต่าง ๆ จากนั้นจึงเลือกทางเลือกที่ดีที่สุดเพียงอย่างเดียว

(9) โดยทั่วไปแล้วแบบจำลองชนิดนี้มักจะนำมาใช้ เพื่อรวบรวมปัญหาของเหตุการณ์จริงที่มีความซับซ้อนกล่าวคือหากเป็นปัญหาง่าย ๆ ก็ไม่จำเป็นต้องใช้แบบจำลองชนิดนี้ ตัวอย่างเช่น อาจใช้แบบจำลองสถานการณ์เพื่อการกระจายความน่าจะเป็นจริง ๆ มากกว่าการนำมาใช้เพื่อการประมาณการธรรมดาซึ่งการกระจายนั้นเรากระจายโดยนำทฤษฎีมาใช้ด้วย

(10) สามารถใช้แบบจำลองสถานการณ์ เพื่อเป็นเครื่องมือวัดประสิทธิภาพของตัวแปรได้ง่ายมากและยังสามารถสะท้อนกลับมาถึงผู้ตัดสินใจได้โดยตรง

2.2 งานวิจัยที่เกี่ยวข้อง

Buck (1960) (อ้างถึงใน Little & Rubin, 1987) เสนอเงื่อนไขของวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งอยู่ในรูปของวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย โดยสร้างสมการถดถอยระหว่างตัวแปรที่สนใจ y กับตัวแปรช่วย x โดยใช้ข้อมูลที่สมบูรณ์ทั้งหมด r หน่วยแล้วประมาณค่าข้อมูลสูญหายด้วยสมการถดถอย $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ นำค่า $\hat{y}_i$ ที่ได้ไปแทนที่ข้อมูลสูญหายของตัวแปรที่สนใจ y โดยกำหนดให้

$$y_{\cdot i} = \begin{cases} y_i & if\ i \in R \\ \hat{\beta}_0 + \hat{\beta}_1 x_i & if\ i \in R^c \end{cases} \tag{2.9}$$

เมื่อ

$y_{\cdot i}$ คือ ค่าของตัวแปรที่สนใจ y ของข้อมูลที่สมบูรณ์ r หน่วย และค่าของตัวแปรที่สนใจ y ของหน่วยสังเกตุที่ i หลังทำการประมาณค่าข้อมูลสูญหาย $n-r$ หน่วย

y_i คือ ค่าของตัวแปรที่สนใจ y ของหน่วยสังเกตุที่ i ซึ่งมีทั้งหมด r หน่วย

x_i คือ ค่าของตัวแปรช่วย x ของหน่วยสังเกตุที่ i โดยที่ค่าของตัวแปรที่สนใจ y เป็นค่าข้อมูลสูญหาย

Lee, Rancourt, and Sarndal (1994) ศึกษาวิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วนและวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด ซึ่งวิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วนเป็นวิธีการประมาณค่าพารามิเตอร์เดียวซึ่งศึกษาประชากรจำกัดขนาด N โดยที่ r เป็นจำนวนหน่วยของข้อมูลที่สมบูรณ์ โดยการชักตัวอย่างแบบสุ่มเชิงเดียวชนิดไม่คืนที่ (Simple random sampling without replacement หรือ *SRSWOR*) จากตัวอย่างที่สุ่มมา s หน่วย จากประชากร $N = 100$ แล้วตัดข้อมูลโดยสุ่ม จากนั้นแทนค่าข้อมูลสูญหายด้วยค่าของตัวประมาณค่าแบบอัตราส่วนโดยใช้ข้อมูลที่สมบูรณ์ทั้งหมด r หน่วย กำหนดให้

$$y_{\cdot i} = \begin{cases} y_i & if\ i \in R \\ \hat{b}x_i & if\ i \in R^c \end{cases} \quad \text{เมื่อ} \quad \hat{b} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_i} \tag{2.10}$$

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.10) อยู่ในรูป

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{2.11}$$

เมื่อ

$\bar{x}_r$ คือ ค่าเฉลี่ยของตัวแปรช่วย x ที่มีค่าข้อมูลของตัวแปรที่สนใจ y สมบูรณ์ทั้งหมด r หน่วย

$\bar{x}_n$ คือ ค่าเฉลี่ยของตัวแปรช่วย x จากข้อมูลที่สมบูรณ์ทั้งหมด n หน่วย

$\bar{y}_r$ คือ ค่าเฉลี่ยของตัวแปรที่สนใจ y จากข้อมูลที่สมบูรณ์ทั้งหมด r หน่วย เป็นตัวประมาณค่าที่เอนเอียง โดยมีความเอนเอียง (biasedness) ดังนี้

$$b(\bar{y}_{RAT}) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}(C_X^2 - \rho_{YX}C_Y C_X) \tag{2.12}$$

และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$MSE(\bar{y}_{RAT}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left\{S_Y^2 + \left(\frac{\bar{Y}}{\bar{X}}\right)^2 S_X^2 - 2\left(\frac{\bar{Y}}{\bar{X}}\right)S_{XY}\right\} \tag{2.13}$$

ส่วนวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด เป็นวิธีการประมาณค่าพารามิเตอร์เดียว กำหนดให้

$$y_{\cdot_i} = \begin{cases} y_i & if\ i \in R \\ y_{NNi} & if\ i \in R^c \end{cases} \tag{2.14}$$

เมื่อ $y_{NNi} = y_l$ ถ้า $\min_{1<l<r}|x_i - x_l|$ สำหรับ $l \in r$ และ $l \neq i$

และ y_{NNi} คือ ค่าของตัวแปรที่สนใจ y เป็นค่าข้อมูลสูญหายที่ถูกประมาณค่าด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด

นอกจากนี้พิจารณาเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายทั้ง 3 วิธี คือวิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน วิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุดและวิธีการประมาณข้อมูลสูญหายด้วยค่ามัลติเพิล (Multiple imputation) โดยใช้เทคนิคมอนติคาร์โล ชักตัวอย่างขนาด 30% ของขนาดประชากรและชักตัวอย่างด้วยอัตราค่าข้อมูลสูญหาย 30% ของขนาดตัวอย่าง ซึ่งค่าของตัวแปรช่วย x มีค่าเฉลี่ยเท่ากับ 48 และค่าความแปรปรวนเท่ากับ 768 และมีการแจกแจงแบบแกมมา (Gamma- distribution) โดยที่ตัวแปรช่วย x มีความสัมพันธ์ในระดับ 0.75 กับตัวแปรที่สนใจ y ที่มีการแจกแจงแบบแกมมา เช่นกันและมีค่าเฉลี่ยเท่ากับ $\mu(x) = a + bx + cx^2$ และค่าความแปรปรวนเท่ากับ $\sigma^2(x) = d^2x^{2g}$ กำหนดให้ค่า a, b, c, d และ g คือค่าคงตัว สามารถเขียนการแจกแจงแบบแกมมาได้ดังนี้ $\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp(-x/\beta)$ โดยที่ $x > 0$ และค่า $\mu(x) = a + bx + cx^2 = \alpha\beta$ และค่า $\sigma^2(x) = d^2x^{2g} = \alpha\beta^2$ โดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบ 3 วิธี คือรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียงสัมพัทธ์สัมบูรณ์ (Absolute relative bias) และอัตราความครอบคลุม (Coverage rate) ได้ข้อสรุปว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วนเป็นวิธีที่เหมาะสมเมื่อพิจารณาจากเกณฑ์ของวิธีค่าความเอนเอียงสัมพัทธ์สัมบูรณ์

Chaimongkol and Suwattee (2005 c) ศึกษาวิจัยเรื่อง Weighted Nearest Neighbor and Regression Imputation ซึ่งเป็นวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุดและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยมารวมกัน เพื่อนำมาใช้ในการจัดการกับข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามหรือเฉพาะบางตัวแปร โดยมีข้อสมมุติว่าความแปรผันของ x และ y มีความสัมพันธ์กันในรูปแบบ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ โดยที่ β_0 และ β_1 คือค่าคงตัวและ $\varepsilon_i \sim NID(0,\sigma^2)$ โดยกำหนดให้

$$y_{\cdot i} = \begin{cases} y_i & if\ i \in R \\ \hat{y}_i^{\bullet} = w_i y_i^{\bullet} + (1-w_i)\hat{y}_i & if\ i \in R^c \end{cases} \tag{2.15}$$

เมื่อ $y_i^{\bullet}$ คือ การประมาณค่าข้อมูลสูญหายของตัวแปรที่สนใจ y ด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด

$\hat{y}_i$ คือ การประมาณค่าข้อมูลสูญหายของตัวแปรที่สนใจ y ด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย

และ $w_i = \dfrac{\sum_{i=1}^{r}(x_i-\bar{x})^2 + r(x_i-\bar{x})^2}{(r+1)\sum_{i=1}^{r}(x_i-\bar{x})^2 + r(x_i-\bar{x})^2}$ เป็นค่าที่ใช้ในการถ่วงน้ำหนัก

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.15) อยู่ในรูป

$$\bar{y}_{WNR} = \frac{1}{n}\sum_{i=1}^{n} y_{\cdot i} \tag{2.16}$$

และมีค่าความแปรปรวน

$$V(\bar{y}_{WNR}) = \sigma^2 \left\{ \frac{\sum_{i=1}^{r}(x_i-\bar{x})^2 + r(x_i-\bar{x})^2}{(r+1)\sum_{i=1}^{r}(x_i-\bar{x})^2 + r(x_i-\bar{x})^2} \right\} \tag{2.17}$$

โดยที่

$$\sigma^2 = \left\{ \frac{\sum_{i=1}^{r}(y_i-\hat{y}_i^{\bullet})^2}{(r-1)\bar{x}\left\{1-\left(\frac{1}{r}\right)(CV_{Xr})^2\right\}} \right\} \tag{2.18}$$

และ $CV_{Xr} = \dfrac{S_{Xr}}{\bar{x}_r}$

ในงานวิจัยนี้พิจารณาเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายทั้ง 3 วิธี คือ วิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย และวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนักโดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบ 4 วิธีคือรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียงสัมพัทธ์สัมบูรณ์ ค่าอัตราส่วนความสำเร็จในการประมาณค่า (the ratio for the performances of the imputations) และอัตราความครอบคลุม โดยใช้เทคนิคมอนติคาร์โลจำลองข้อมูลประชากรจำกัดขนาด N = 400 สุ่มตัวอย่างขนาด n มาจำนวน10%, 20% และ 30% ของขนาดประชากร ด้วยวิธีการชักตัวอย่างแบบง่ายไม่แทนที่ อัตราค่าของข้อมูลสูญหายที่คาดหวัง E(m/n) ที่ใช้ในการพิจารณา คือ 10%, 20% และ30%โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร x และ y คือ 0.882 ได้ข้อสรุปว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนักเป็นวิธีที่เหมาะสมเมื่อพิจารณาจากเกณฑ์ของวิธีรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย

Chaimongkol and Suwattee (2005 a) ศึกษาวิจัยเรื่อง The Effect of Weighted Nearest Neighbor–Regression Imputation on Simple Linear Regression Analysis เปรียบเทียบผลกระทบของวิธีการประมาณค่าข้อมูลสูญหาย 4 วิธี คือ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยโดยสุ่ม วิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค และวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก ในการวิเคราะห์ถดถอยอย่างง่ายโดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบ 4 วิธี คือ ค่าสัมประสิทธิ์สหสัมพันธ์ ค่าสัมประสิทธิ์การตัดสินใจ ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean absolute deviation) และช่วงความเชื่อมั่นของค่าเฉลี่ย ซึ่งมีอัตราค่าของข้อมูลสูญหายที่คาดหวังที่ใช้ในการพิจารณาคือ 10% 15% และ 20% สรุปว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก เป็นวิธีที่เหมาะสม เมื่อพิจารณาจากเกณฑ์ของค่าสัมประสิทธิ์สหสัมพันธ์ ค่าสัมประสิทธิ์การตัดสินใจดีกว่าวิธีอื่น ๆ ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดและมีช่วงความเชื่อมั่นของค่าเฉลี่ยแคบกว่าวิธีอื่น ๆ ภายใต้ตัวแบบจำลองที่ศึกษา

Singh et al. (2010) ศึกษาวิจัยเรื่อง Estimation of population mean using imputation techniques in sample surveys โดยเสนอตัวประมาณค่าข้อมูลสูญหาย 2 วิธี วิธีแรกคือการแทนค่าข้อมูลสูญหาย โดยใช้เฉพาะข้อมูลของตัวแปรช่วย x ของหน่วยตัวอย่างที่ข้อมูลของตัวแปรที่สนใจ y สมบูรณ์ซึ่งมีทั้งหมด r หน่วย $(r < n)$ โดยกำหนดให้

$$y_{\cdot i} = \begin{cases} y_i & if\ i \in R \\ \dfrac{1}{(n-r)}\bar{y}_r \left[\dfrac{n\{(A+C)\bar{X} + fB\bar{x}_r\}}{\{(A+fB)\bar{X} + C\bar{x}_r\}} - r \right] & if\ i \in R^c \end{cases} \qquad (2.19)$$

เมื่อ $A=(d-1)(d-2)$, $B=(d-1)(d-4)$, $C=(d-2)(d-3)(d-4)$ และ $f=\frac{n}{N}$

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.19) อยู่ในรูป

$$T_{d1}=\bar{y}_r\left[\frac{(A+C)\bar{X}+fB\bar{x}_r}{(A+fB)\bar{X}+C\bar{x}_r}\right] \tag{2.20}$$

เป็นตัวประมาณค่าที่เอนเอียง โดยมีความเอนเอียง (biasedness) ดังนี้

$$b(T_{d1})=\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}\varphi\left[\rho_{YX}C_YC_X-\varphi_2C_X^2\right] \tag{2.21}$$

เมื่อ $\varphi=\varphi_1-\varphi_2$, $\varphi_1=\frac{fB}{A+fB+C}$ และ $\varphi_2=\frac{C}{A+fB+C}$

โดยที่ C_Y คือ สัมประสิทธิ์การแปรผันของข้อมูลของตัวแปรที่สนใจ Y

C_X คือ สัมประสิทธิ์การแปรผันของข้อมูลของตัวแปรช่วย X

และ ρ_{YX} คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรที่สนใจ Y และตัวแปรช่วย X

และมีค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$MSE(T_{d1})=\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2\left[C_Y^2+\varphi^2C_X^2+2\varphi\rho_{YX}C_YC_X\right] \tag{2.22}$$

ซึ่งพบว่า $MSE(T_{d1})$ มีค่าต่ำสุดก็ต่อเมื่อ

$$\varphi=\left[\frac{fB-C}{A+fB+C}\right]=-\rho_{YX}\frac{C_Y}{C_X} \tag{2.23}$$

จากสมการ (2.23) เมื่อทราบค่า ρ_{YX} C_Y และ C_X สามารถหาค่า d ที่เหมาะสมได้ และจะได้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่เหมาะสมดังนี้

$$MSE(T_{d1})_{opt}=\left(\frac{1}{r}-\frac{1}{N}\right)(1-\rho_{YX}^2)S_Y^2 \tag{2.24}$$

เมื่อ S_Y^2 คือ ความแปรปรวนของข้อมูลของตัวแปรที่สนใจ Y

สำหรับวิธีที่สองที่ Singh et al. เสนอ คือการแทนค่าข้อมูลสูญหายโดยใช้ข้อมูลของตัวแปรช่วย x จากทั้ง n หน่วย โดยกำหนดให้

$$y_{*_i} = \begin{cases} y_i & if\ i \in R \\ \dfrac{1}{(n-r)}\bar{y}_r \left[\dfrac{n\{(A+C)\bar{X} + fB\bar{x}_n\}}{\{(A+fB)\bar{X} + C\bar{x}_n\}} - r \right] & if\ i \in R^c \end{cases} \quad (2.25)$$

ตัวประมาณค่าเฉลี่ยประชากรที่ใช้ข้อมูลที่ประมาณค่าข้อมูลสูญหายด้วยสมการ (2.25) อยู่ในรูป

$$T_{d2} = \bar{y}_r \left[\frac{(A+C)\bar{X} + fB\bar{x}_n}{(A+fB)\bar{X} + C\bar{x}_n} \right] \quad (2.26)$$

ความเอนเอียง (biasedness) คือ

$$b(T_{d2}) = \left(\frac{1}{n} - \frac{1}{N} \right) \bar{Y} \varphi \left[\rho_{YX} C_Y C_X - \varphi_2 C_X^2 \right] \quad (2.27)$$

มีค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$MSE(T_{d2}) = \bar{Y}^2 \left[\left(\frac{1}{r} - \frac{1}{N} \right) C_Y^2 + \left(\frac{1}{n} - \frac{1}{N} \right) \left[\varphi^2 C_X^2 + 2\varphi \rho_{YX} C_Y C_X \right] \right] \quad (2.28)$$

และค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่เหมาะสมดังนี้

$$MSE(T_{d2})_{opt} = \left[\left(\frac{1}{r} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{N} \right) \rho_{YX}^2 \right] S_Y^2 \quad (2.29)$$

พิจารณาเปรียบเทียบประสิทธิภาพของตัวประมาณค่าที่ใช้ประมาณค่า $\bar{Y}$ ได้แก่ ตัวประมาณค่าที่ใช้วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วนและวิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ ตัวประมาณค่า T_{d1} และ T_{d2} โดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบ คือ ประสิทธิภาพสัมพัทธ์ (Relative efficiency) ได้ข้อสรุปว่าตัวประมาณค่า T_{d1} เป็นวิธีที่ดีที่สุดสามารถอธิบายได้อย่างมีประสิทธิภาพเมื่อศึกษาด้วยการจำลอง

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องจะเห็นได้ว่าตัวประมาณค่า T_{d1} มีประสิทธิภาพดีที่สุดเมื่อเปรียบเทียบกับวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน วิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ และตัวประมาณค่า T_{d2} ผู้วิจัยจึงสนใจที่จะพัฒนาตัวประมาณค่า T_{d1} ให้มีประสิทธิภาพสูงยิ่งขึ้น