

เปิดประเด็น

Issues Raising

สวະโชติ ศรีสุทธียากร

การวิเคราะห์ข้อมูลสูญหาย

Missing data analysis

การเก็บรวบรวมข้อมูลในงานวิจัยทางด้านสังคมศาสตร์นั้นมักเกิดปัญหาข้อมูลสูญหาย (missing data) กล่าวคือบางหน่วยตัวอย่างไม่ตอบข้อคำถาม (unit nonresponse) หรือเกิดการไม่ตอบในเฉพาะบางคำถาม (item nonresponse) ข้อมูลสูญหายที่เกิดขึ้นอาจส่งผลกระทบต่อ การอนุมานเชิงสถิติ เช่น มีผลกระทบที่มีต่อค่าประมาณพารามิเตอร์ (parameter estimated value) ปัญหาข้อมูลสูญหายอาจทำให้ตัวประมาณค่าพารามิเตอร์เป็นตัวประมาณที่มีความลำเอียง (biased) (Rubin, 1987) ยกตัวอย่างเช่น ในการสำรวจรายได้ส่วนบุคคลของประชาชน ในกรุงเทพมหานครอาจมีความเป็นไปได้ที่กลุ่มประชากรที่มีรายได้สูงมีแนวโน้มเลือกที่จะไม่ตอบข้อมูลรายได้ส่วนบุคคลมากกว่ากลุ่มประชากรที่มีรายได้ปานกลางและรายได้ต่ำ ในกรณีนี้ กลุ่มตัวอย่างที่นักวิจัยมีจึงเป็นกลุ่มตัวอย่างที่ไม่สามารถเป็นตัวแทนของประชากรที่ดีได้ ค่าประมาณรายได้ของประชากรที่ได้จากตัวอย่างนี้จึงเป็นค่าประมาณที่มีความลำเอียง หรือ มีผลต่ออำนาจการทดสอบสมมติฐานทางสถิติ (statistical power) กล่าวคือ เมื่อเกิดปัญหาข้อมูลสูญหายและนักวิจัยเลือกใช้การแก้ปัญหาด้วยวิธีการตัดข้อมูล องศาความเป็นอิสระ (degree of freedom) ของสถิติทดสอบมีค่าน้อยลงในขณะที่ค่าคลาดเคลื่อนมาตรฐาน (standard error) ของค่าสถิติจะมีค่าสูงขึ้น ส่งผลให้ อำนาจในการทดสอบสมมติฐานทางสถิติลดลง เมื่อเปรียบเทียบกับสถานการณ์ที่ผู้วิจัยมีข้อมูลครบถ้วนสมบูรณ์ (Cohen & Cohen, 1983; Cool, 2000)

ความถูกต้องของผลการวิเคราะห์เมื่อเกิดข้อมูลสูญหายนั้นขึ้นกับปัจจัยหลักหลาย ปัจจัยได้แก่ ขนาดการสูญหายของข้อมูล (size of missing data) รูปแบบการสูญหายของ ข้อมูล (pattern of missing data) และวิธีการแก้ไขปัญหามissing data ของข้อมูล (method for treating missing data) ผู้เขียนได้แบ่งบทความนี้ออกเป็น ๔ ส่วน ได้แก่ (๑) ปัญหาที่เกิด จากข้อมูลสูญหาย (problem caused by missing data) (๒) รูปแบบการสูญหายของข้อมูล (pattern of missing data) (๓) ข้อเสนอแนะในการจัดการกับข้อมูลสูญหาย

๑. รูปแบบการสูญหายของข้อมูล (pattern of missing data)

รูปแบบการสูญหายของข้อมูลเป็นปัจจัยหลักที่ส่งอิทธิพลต่อผลการวิเคราะห์มากกว่าขนาดการสูญหายของข้อมูล (Tabachnick & Fidell, 2001) นักวิจัยจำเป็นต้องทราบรูปแบบการสูญหายของข้อมูลเพื่อใช้ในการเลือกวิธีการแก้ปัญหาที่เหมาะสม รูปแบบหรือกลไกการสูญหายของข้อมูลนั้นสามารถแบ่งออกได้เป็น ๓ ประเภทหลัก (Little & Rubin, 1987) ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (missing completely at random) การสูญหายแบบสุ่ม (missing at random) และการสูญหายแบบไม่สุ่ม (missing not at random) ซึ่งมีรายละเอียดต่อไปนี้ สมมติให้ข้อมูลตัวอย่างจากการวิจัยเชิงสำรวจประกอบด้วยตัวแปรรายได้ต่อเดือน (Y) เพศ (X1) และระดับการศึกษา (X2)

๑.๑ การสูญหายแบบสุ่มสมบูรณ์ (missing completely at random: MCAR)

เป็นรูปแบบการสูญหายที่เป็นรูปแบบสุ่มโดยไม่มีความสัมพันธ์กับตัวแปรใดๆหรือค่าข้อมูลอื่นๆ ทั้งที่ทราบและไม่ทราบค่า จากตัวอย่างหากตัวแปรรายได้ต่อเดือน (Y) เกิดการสูญหายแบบสุ่มสมบูรณ์ การเกิดการสูญหายในตัวแปร Y จะไม่ได้มีความสัมพันธ์กับตัวแปร X1 และ X2 รวมทั้งไม่ได้มีความสัมพันธ์กับการสูญหายในตัวแปร Y เองด้วย เนื่องจากการสูญหายในลักษณะนี้เกิดขึ้นอย่างสุ่มโดยไม่ขึ้นกับปัจจัยใดๆ ดังนั้นข้อมูลของตัวแปรที่มีการสูญหายจึงเทียบเท่ากับตัวอย่างสุ่มจากตัวอย่างข้อมูลที่สมบูรณ์หรือไม่มีการสูญหาย ดังนั้นหากเกิดการสูญหายในรูปแบบนี้นักวิจัยสามารถละทิ้งหน่วยตัวอย่างที่มีข้อมูลสูญหายได้โดยที่ไม่ส่งผลกระทบทำให้เกิดความลำเอียงในการประมาณค่าพารามิเตอร์ แต่อาจลดทอนอำนาจในการทดสอบสมมติฐานทางสถิติเมื่อเปรียบเทียบกับการใช้ข้อมูลสมบูรณ์

๑.๒ การสูญหายแบบสุ่ม (missing at random: MAR)

อาจเรียกว่าการสูญหายแบบสุ่มที่มีเงื่อนไข (missing conditionally at random) เป็นการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มโดยสมบูรณ์ทั้งชุดข้อมูลค่าสังเกตแต่เกิดขึ้นอย่างสุ่มเพียงบางส่วนหรือบางกลุ่มของค่าสังเกต กล่าวคือการสูญหายของค่าสังเกตมีความสัมพันธ์กับตัวแปรหรือค่าข้อมูลอื่นๆ ที่ทราบค่าแต่ไม่มีความสัมพันธ์กับข้อมูลที่ไม่ทราบค่า จากตัวอย่าง หากเกิดการสูญหายของค่าสังเกตตัวแปร Y โดยที่พบว่าการสูญหายนั้นมีความสัมพันธ์กับ X1 หรือ X2 เท่านั้น จะกล่าวว่าการสูญหายที่เกิดขึ้นในตัวแปรรายได้ต่อเดือนนี้เป็นการสูญหายแบบสุ่ม (MAR) รูปแบบการสูญหายแบบสุ่มสามารถตรวจสอบได้จากข้อมูล ผู้วิจัยอาจใช้การวิเคราะห์ความถดถอยโลจิสติก (logistic regression) โดยที่กำหนดให้ตัวแปรตามคือการเกิดการสูญหายในตัวแปรรายได้ต่อเดือนโดยมีค่าเท่ากับ ๑ ถ้าข้อมูลไม่สูญหาย และมีค่าเท่ากับ ๐ ถ้าข้อมูลสูญหาย และตัวแปรอิสระคือเพศและระดับการศึกษา

๑.๓ การสูญหายแบบไม่สุ่ม (missing not at random: MNAR) เป็นรูปแบบการสูญหายที่การสูญหายข้อมูลของตัวแปร Y มีความสัมพันธ์กับตัวแปร Y เองและกับตัวแปร X1 หรือ X2 ยกตัวอย่างเช่น หากหน่วยตัวอย่างเพศหญิงที่มีรายได้ต่อเดือนอยู่ในระดับสูงเลือกจะไม่ตอบคำถามรายได้ต่อเดือนมากกว่าเพศชาย การสูญหายแบบ MNAR ก่อให้เกิดปัญหาในการวิเคราะห์ที่ร้ายแรงมากกว่า MCAR และ MAR เพราะค่าประมาณพารามิเตอร์ที่ได้จะมีความลำเอียงทำให้เกิดความคลาดเคลื่อนของผลการวิจัย

๒. วิธีการแก้ไขปัญหาคข้อมูลสูญหาย (missing data treatments)

ในปัจจุบันได้มีการพัฒนาวิธีการเพื่อแก้ไขปัญหาคข้อมูลสูญหายไว้หลากหลายวิธีการ ซึ่งมีรายละเอียดดังต่อไปนี้

วิธีการตัดข้อมูล (deletion method) วิธีการนี้ใช้การตัดข้อมูลที่สูญหายออกจากการวิเคราะห์ ซึ่งการตัดข้อมูลสูญหายสามารถกระทำได้สองลักษณะ ได้แก่ (๑) การตัดข้อมูลสูญหายแบบ Listwise (listwise deletion or complete case analysis) วิธีการนี้กระทำได้ง่ายโดยการตัดข้อมูลหน่วยตัวอย่างที่มีการสูญหายของข้อมูลออกจากการวิเคราะห์ทั้งหมด จากนั้นจึงนำข้อมูลส่วนที่เหลือไปวิเคราะห์ต่อไป ถึงแม้ว่าวิธีการนี้จะกระทำได้ง่ายและสะดวกแต่มีข้อควรระวังคือวิธีการนี้ไม่สามารถใช้ได้ในทุกกรณีกล่าวคือ หากกลไกการสูญหายของข้อมูลเป็นแบบ MCAR วิธีการตัดข้อมูลแบบ Listwise จะสามารถกระทำได้เพราะค่าประมาณพารามิเตอร์ที่ได้จะมีความไม่ลำเอียง (unbiased) แต่ค่าคลาดเคลื่อนมาตรฐานของค่าประมาณพารามิเตอร์อาจมีค่าสูงขึ้นเนื่องจากขนาดตัวอย่างมีน้อยลง ทำให้ลดทอนอำนาจของการทดสอบสมมติฐานลง แต่อย่างไรก็ตามอำนาจการทดสอบสมมติฐานมักไม่ได้รับผลกระทบมากนักหากขนาดตัวอย่างมีขนาดใหญ่ แต่หากกลไกการสูญหายของข้อมูลไม่เป็นแบบ MCAR การใช้วิธีการตัดข้อมูลแบบ Listwise จะไม่ใช่วิธีการที่เหมาะสมที่สุด (๒) การตัดข้อมูลสูญหายแบบ pairwise (pairwise deletion or available case analysis) วิธีการนี้ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ เช่น ค่าสัมประสิทธิ์สหสัมพันธ์ โดยจะไม่ตัดข้อมูลของหน่วยตัวอย่างที่มีข้อมูลสูญหายออกทั้งหมดแต่จะทำการวิเคราะห์ข้อมูลเฉพาะข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปรแทน วิธีการนี้มักไม่เป็นที่ยอมรับใช้ เหตุผลหนึ่งคือบ่อยครั้งที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรในชุดข้อมูลนั้นคำนวณจากตัวอย่างคนละชุดกัน ซึ่งจะทำให้การประมาณค่าพารามิเตอร์ในโมเดลที่ต้องใช้เมทริกซ์สหสัมพันธ์เป็นข้อมูลในการคำนวณ เช่น โมเดลสมการโครงสร้าง (structural equation model) อาจไม่สามารถหาผลเฉลยได้เพราะเมทริกซ์สหสัมพันธ์ที่สร้างขึ้นจากตัวอย่างคนละชุดกันนั้นอาจหาตัวผกผัน (invert) ไม่ได้

วิธีการปรับค่าโดยใช้ตัวแปรหุ่น (dummy variable adjustment) วิธีการนี้เป็นวิธีการที่ใช้แก้ไขปัญหาค่าสังเกตสูญหายในตัวแปรอิสระในการวิเคราะห์ความถดถอย (regression analysis) หลักการของวิธีการนี้สามารถอธิบายได้ดังนี้ สมมติว่า Y เป็นตัวแปรตาม และตัวแปร X เป็นตัวแปรอิสระตัวหนึ่งในชุดข้อมูลที่มีข้อมูลสูญหาย การแก้ไขปัญหามิสามารถทำได้โดยการสร้างตัวแปรหุ่น (dummy variable: D) โดยกำหนดให้ $D = 1$ ถ้าค่าสังเกต X สูญหาย และ $D = 0$ ถ้าค่าสังเกต X ไม่สูญหาย และสร้างตัวแปร X^* โดยที่ $X^* = X$ ถ้าค่าสังเกต X ไม่มีการสูญหาย และ $X^* = c$ ถ้าค่าสังเกต X มีการสูญหาย โดยที่ c คือค่าคงที่ซึ่งอาจเป็นค่าเฉลี่ยหรือฐานนิยม เป็นต้น จากนั้นวิเคราะห์สมการถดถอยโดยกำหนดให้ Y เป็นตัวแปรตาม และตัวแปรอิสระคือ X^* , D และตัวแปรอื่นที่เหลือ ค่าสัมประสิทธิ์ความถดถอยของตัวแปร X ที่มีค่าสูญหายจะถูกปรับด้วยสัมประสิทธิ์ความถดถอยของตัวแปรหุ่น D อย่างไรก็ตามจากงานวิจัยของ Jones (1996) พบว่าวิธีการนี้จะก่อให้เกิดความลำเอียงในการประมาณค่าสัมประสิทธิ์ความถดถอย

วิธีการทดแทนค่าสูญหาย (imputation method) วิธีการนี้อาจแบ่งออกได้เป็น ๒ วิธีการ ได้แก่ (๑) วิธีการด้วยค่าเฉลี่ย (mean imputation method) มีหลักการคือจะทดแทนค่าสังเกตของตัวแปรที่สูญหายด้วยค่าเฉลี่ยของตัวอย่างที่มีข้อมูลสมบูรณ์ แล้วจึงนำข้อมูลที่ผ่านการทดแทนแล้วไปทำการวิเคราะห์ต่อไป นักวิจัยบางส่วนนิยมใช้วิธีการนี้ อย่างไรก็ตามวิธีการนี้อาจไม่เหมาะสมถ้ากลไกการสูญหายไม่ได้เป็นแบบ MCAR เพราะค่าประมาณพารามิเตอร์ที่ได้จะมีความลำเอียง (biased) นอกจากนี้การทดแทนค่าสูญหายด้วยค่าเฉลี่ยหรือฐานนิยมยังทำให้การประมาณค่าความแปรปรวน (variance) ความแปรปรวนร่วม (covariance) และค่าคลาดเคลื่อนมาตรฐานของค่าประมาณพารามิเตอร์มีความลำเอียง ซึ่งจะทำให้การอนุมานเชิงสถิติเกี่ยวกับพารามิเตอร์ในโมเดลนั้นมีความคลาดเคลื่อนไม่น่าเชื่อถือ (๒) วิธีการแทนค่าด้วยสมการถดถอย (regression-based imputation) เป็นวิธีการที่เหมาะสมกับการสูญหายแบบ MAR สามารถทำได้โดยโมดูล missing value analysis วิธีการนี้ใช้สมการถดถอย (regression equation) เป็นเครื่องมือในการทำนายค่าสูญหายของตัวแปรที่สนใจโดยใช้ตัวแปรอื่นที่มีในชุดข้อมูล อย่างไรก็ตามวิธีการนี้จะให้ค่าคลาดเคลื่อนมาตรฐานต่ำกว่าปกติ ซึ่งสามารถแก้ไขได้โดยการเพิ่มพจน์ของความคลาดเคลื่อนสุ่มที่มีความแปรปรวนเท่ากับค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (mean square error: MSE) เข้าไปในสมการถดถอยซึ่งจะทำให้ค่าคลาดเคลื่อนมาตรฐานมีค่าใกล้เคียงค่าจริงมากขึ้น (๓) วิธีการค่าคาดหวังสูงสุด (expectation - maximization method) วิธีการนี้มักใช้อัลกอริทึม EM (EM algorithm) เครื่องมือในการคำนวณ สามารถทำได้โดยใช้โมดูล missing value analysis เช่นเดียวกัน ซึ่งมีหลักการคือ แบ่งชุดข้อมูลออกเป็นสองส่วน ได้แก่ ส่วนที่ไม่มีข้อมูลสูญหาย (nonmissing) และส่วนที่มีข้อมูลสูญหาย (missing) และกำหนดค่าเริ่มต้นให้กับพารามิเตอร์ที่ต้องการประมาณ จากนั้นในขั้นตอนแรกของอัลกอริทึม

(E step) จะใช้ค่าพารามิเตอร์เพื่อประมาณค่าสูญหายของข้อมูล จากนั้นขั้นตอนที่สอง (M step) จะใช้ข้อมูลที่ได้ทดแทนค่าสูญหายแล้วในการประมาณค่าพารามิเตอร์ที่ต้องการด้วยการใช้วิธีภาวะความควรจะเป็นสูงสุด (maximum likelihood) ทวนซ้ำอัลกอริทึมทั้งสองขั้นตอนจนกว่าค่าประมาณที่ได้จากลู่อู่เข้าจึงหยุดอัลกอริทึม วิธีการนี้เหมาะกับข้อมูลสูญหายแบบ MAR โปรแกรม SPSS (missing value analysis) สามารถทดแทนค่าสูญหายด้วยวิธีการนี้ได้ ค่าประมาณพารามิเตอร์ที่ได้จากวิธีการนี้จะมีคุณสมบัติที่ดีคือ มีความไม่ลำเอียงเมื่อเข้าใกล้อนันต์ (asymptotically unbiased) แต่อย่างไรก็ตามค่าประมาณจากวิธีการนี้มีแนวโน้มที่จะมีค่าคลาดเคลื่อนมาตรฐานที่ต่ำกว่าปกติ และ (๔) วิธีการทดแทนหลายค่า (multiple imputation) วิธีการนี้พัฒนาเพื่อลดข้อด้อยของวิธีการทดแทนค่าเดียวในช่วงต้น สามารถกระทำได้โดยใช้โมดูล multiple imputation วิธีการนี้ประกอบไปด้วย ๓ ขั้นตอนได้แก่ (๑) สร้างข้อมูลสมบูรณ์โดยใช้หลักการทดแทนค่าขึ้นมาจำนวนหนึ่ง โดยทั่วไปนิยมสร้างประมาณ ๕-๑๐ ชุดข้อมูล (๒) ประมาณค่าพารามิเตอร์ที่ต้องการจากแต่ละชุดข้อมูล ในขั้นตอนนี้จะมีการคำนวณค่าพารามิเตอร์ที่สนใจเท่ากับจำนวนชุดข้อมูลที่สร้างขึ้นในขั้นตอนแรก (๓) จากนั้นนำค่าประมาณพารามิเตอร์และค่าคลาดเคลื่อนมาตรฐานที่คำนวณได้จากแต่ละชุดข้อมูลมารวมกันด้วยการเฉลี่ย วิธีการนี้สามารถลดทอนปริมาณความลำเอียงในการประมาณค่าคลาดเคลื่อนมาตรฐานได้ดีกว่าการทดแทนค่าเดียว เพราะมีการรวมความผันแปรในการทดแทนข้อมูลไว้ในในการประมาณค่าพารามิเตอร์ จากผลการศึกษาของ Schafer (1997) พบว่า เมื่อการสูญหายของข้อมูลเป็นแบบ MAR โดยมีค่าสูญหายร้อยละ ๓๐ ของข้อมูลทั้งหมด วิธีการทดแทนหลายค่าโดยใช้ชุดข้อมูล ๕ ชุด จะมีประสิทธิภาพคิดเป็นร้อยละ ๙๔ เมื่อเปรียบเทียบกับสถานการณ์ที่ไม่มีข้อมูลสูญหาย และเมื่อมีค่าสูญหายเกิดขึ้นร้อยละ ๕๐ วิธีการทดแทนหลายค่าโดยใช้ชุดข้อมูล ๑๐ ชุด จะมีประสิทธิภาพคิดเป็นร้อยละ ๙๔ เช่นเดียวกันเมื่อเปรียบเทียบกับสถานการณ์ที่ไม่มีข้อมูลสูญหาย

วิธีการใช้โมเดลการคัดเลือกของ Heckman (Heckman selection model) วิธีการนี้ประกอบไปด้วย ๓ ขั้นตอน ขั้นตอนแรกคือ การประมาณค่าสมการการคัดเลือก (selection equation) ขั้นตอนที่สองคือการสร้างตัวแปรใหม่ที่ไม่มีค่าสูญหาย และขั้นตอนที่สามคือ นำตัวแปรที่สร้างใหม่ไปใช้ในการวิเคราะห์ต่อไป วิธีการนี้เป็นวิธีการที่เหมาะสมกับการแก้ไขปัญหาข้อมูลสูญหายที่มีกลไกการสูญหายแบบ MNAR

๓. ข้อเสนอแนะในการจัดการกับข้อมูลสูญหาย

เมื่อมีข้อมูลสูญหายเกิดขึ้น ขั้นตอนแรกคือการวิเคราะห์จำนวนข้อมูลสูญหายและกลไกการเกิดข้อมูลสูญหาย โดยปกติผู้วิเคราะห์จะไม่สามารถยืนยันได้แน่นอนว่ากลไกการสูญหายของข้อมูลที่เกิดขึ้นนั้นเป็นแบบใด วิธีการหนึ่งซึ่งสามารถใช้ประกอบการพิจารณาได้คือ (๑) พิจารณา

ตัวแปรที่มีข้อมูลสูญหายแล้วสร้างตัวแปรหุ่น (dummy variable: D) โดยกำหนดให้ $D = 1$ ถ้าค่าสังเกต X สูญหาย และ $D = 0$ ถ้าค่าสังเกต X ไม่สูญหาย (๒) วิเคราะห์ความถดถอยโลจิสติก (logistic regression) โดยที่กำหนดให้ตัวแปร D ตัวแปรตาม และตัวแปรอื่นๆ ในชุดข้อมูลเป็นตัวแปรอิสระ (๓) พิจารณาว่ามีตัวแปรใดที่มีอิทธิพลต่อการสูญหายของข้อมูลอย่างมีนัยสำคัญหรือไม่ หากไม่มีกลไกการสูญหายของข้อมูลอาจเป็นแบบ MCAR แต่หากมีกลไกการสูญหายของข้อมูลอาจเป็น MAR เมื่อตรวจสอบกลไกการสูญหายในเบื้องต้นแล้ว ขั้นตอนดำเนินการต่อมาอาจกระทำดังต่อไปนี้

๑. ทดลองใช้วิธีการแก้ปัญหาการสูญหายหลากหลายวิธี เพื่อเปรียบเทียบผลการวิเคราะห์ที่ได้ หากผลการวิเคราะห์ที่ได้มีความสอดคล้องกันแสดงว่าผู้วิเคราะห์สามารถใช้วิธีใดก็ได้ผลไม่แตกต่างกันในกรณีนี้ แต่หากผลการวิเคราะห์ที่ได้มีความแตกต่างกันจะเป็นสิ่งที่บ่งชี้ว่าต้องมีการจัดการปัญหาค่าสูญหายในกรณีนี้อย่างระมัดระวัง

๒. ถ้าการสูญหายที่เกิดขึ้นมีน้อยกว่าร้อยละ ๕ ของขนาดตัวอย่าง วิธีการที่แนะนำคือการใช้การตัดข้อมูลแบบ Listwise การใช้วิธีการทดแทนค่าสูญหายแบบอื่นอาจทำให้เกิดความลำเอียงของผลการวิเคราะห์ได้มากกว่า

๓. ถ้าการสูญหายที่เกิดขึ้นอยู่ระหว่างร้อยละ ๕ ถึงร้อยละ ๑๐ ของขนาดตัวอย่าง ให้พิจารณาว่าการสูญหายนั้นเกิดในตัวแปรอิสระหรือตัวแปรตาม ถ้าการสูญหายที่เกิดขึ้นเกิดในตัวแปรอิสระผู้วิเคราะห์มีทางเลือก ๒ ทาง ได้แก่ (๑) ใช้ตัวแปรอิสระตัวอื่นที่ไม่มีค่าสูญหายแทนตัวแปรอิสระที่มีค่าสูญหาย แต่วิธีการนี้ต้องมั่นใจหรือมีทฤษฎีรองรับว่าตัวแปรทั้งสองสามารถทดแทนกันได้ อย่างไรก็ตามวิธีการนี้กระทำได้ยากในทางปฏิบัติ หรือ (๒) พิจารณาว่าการสูญหายของข้อมูลเป็นแบบ MAR หรือไม่ ถ้าผู้วิเคราะห์เชื่อได้ว่าเป็นการสูญหายแบบ MAR สามารถใช้การตัดข้อมูลแบบ Listwise ได้ ถ้าการสูญหายเกิดขึ้นในตัวแปรตาม ผู้วิเคราะห์สามารถใช้ได้ทั้งวิธีการตัดข้อมูลแบบ Listwise ถ้าเชื่อว่าการสูญหายเป็นแบบ MAR แต่หากการสูญหายไม่เป็นแบบ MAR ให้ใช้โมเดลการคัดเลือกของ Heckman

๔. ถ้าการสูญหายเกิดขึ้นอยู่ระหว่างร้อยละ ๑๐ ถึงร้อยละ ๒๐ ของขนาดตัวอย่าง ผู้วิเคราะห์อาจเลือกใช้วิธีการแก้ปัญหาที่มีประสิทธิภาพมากขึ้น เช่นในกรณีที่มีการสูญหายเป็นแบบ MAR อาจเลือกใช้การทดแทนหลายค่า (multiple imputation) หรือการใช้วิธีค่าคาดหวังสูงสุด (EM method) ในกรณีที่มีการสูญหายไม่เป็นแบบ MAR อาจเลือกใช้โมเดลการคัดเลือกของ Heckman

๕. ถ้าการสูญหายมีมากกว่าร้อยละ ๒๐ ของขนาดตัวอย่าง วิธีการแก้ไขปัญหาค่าสูญหายต่างๆ จะมีประสิทธิภาพที่ต่ำ ผู้วิเคราะห์อาจตัดสินใจตัดตัวแปรที่เกิดการสูญหายนี้ออกจากการวิเคราะห์

รายการอ้างอิง

- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/ correlation analysis for the behavioral sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cool, A. L. (2000). *A review of methods for dealing with missing data*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service No. ED 438 311).
- Jones, M.P. (1996). Indicator and stratication methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91, 222-230.
- Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: Willey.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

ผู้เขียน

อาจารย์ ดร. สวະโชติ ศรีสุทธียากร อาจารย์ประจำ สาขาสถิติการศึกษา ภาควิชาวิจัยและ
จิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร ๑๐๓๓๐
อีเมล: choat.cu@gmail.com