การพัฒนาวิธีการประมาณข้อมูลสูญหายโดยการถ่วงน้ำหนักแบบวนซ้ำ ด้วยวิธีของแจ็คไนฟ์และการวิเคราะห์การถดถอย

The Development of Iterated Weighted Jacknife Method and Regression (IWJR) for Estimating Missing Data

จำลอง วงษ์ประเสริฐ และบุญชม ศรีสะอาด***

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณข้อมูลสูญหายโดยการถ่วงน้ำหนักแบบวนซ้ำด้วยวิธีของแจ็คไนฟ์และการวิเคราะห์การถดถอย (IWJR) และเปรียบเทียบประสิทธิภาพในการประมาณค่าเฉลี่ยประชากร ความแปรปรวนประชากรและสัมประสิทธิ์สหสัมพันธ์ประชากรและอำนาจการทดสอบ ภายใต้ข้อมูลสูญหายแบบสุ่มอย่างสมบูรณ์และใช้การสุ่มตัวอย่างแบบง่าย กับการตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ (LD) วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (MI) และวิธีการประมาณข้อมูลสูญหายด้วยการถดถอย (RI) โดยใช้ข้อมูลจากการจำลองและข้อมูลจริง การเปรียบเทียบกระทำภายใต้เงื่อนไขดังต่อไปนี้ 1) ขนาดตัวอย่าง 3 ขนาด (100 200 และ 500) 2) ระดับความสัมพันธ์ระหว่างตัวแปร 3 ระดับ (ต่ำ ρ = .3 ปานกลาง ρ = .5 และสูง ρ = .7) และ 3) ร้อยละของข้อมูลสูญหาย 4 ระดับ (ร้อยละ 5 10 15 และ 20) และใช้ข้อมูลจากการจำลองศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย ที่ระดับนัยสำคัญ .05

ผลการวิจัยพบว่า

1. ใช้ข้อมูลจากการจำลอง 1) เมื่อจำแนกตามขนาดตัวอย่าง ในการประมาณค่าเฉลี่ยประชากร เมื่อขนาดตัวอย่าง 100 และ 200 RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อขนาดตัวอย่าง 500 LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากรและค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อขนาดตัวอย่าง 100 IWJR มีประสิทธิภาพที่สูงที่สุด เมื่อขนาดตัวอย่าง 200 และ 500 LD และ IWJR มีประสิทธิภาพสูงที่สุด 2) เมื่อจำแนกตามระดับความสัมพันธ์ระหว่างตัวแปร ในการประมาณค่าเฉลี่ยประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ MI RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรปานกลางและสูง RI และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรปานกลาง LD และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรสูง MI RI และ IWJRมีประสิทธิภาพสูงที่สุด 3) เมื่อจำแนกตามร้อยละข้อมูลสูญหาย ในการประมาณค่าเฉลี่ยประชากร เมื่อข้อมูลสูญหายร้อยละ 5 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 10 RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 และร้อยละ 20

** การศึกษาดุษฎีบัณฑิต สาขาวิชาวิจัยและประเมินผลการศึกษา มหาวิทยาลัยมหาสารคาม*

*** รองศาสตราจารย์ ดร., คณะศึกษาศาสตร์ มหาวิทยาลัยมหาสารคาม : ประธานที่ปรึกษาวิทยานิพนธ์*

LD MI RI และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากร เมื่อข้อมูลสูญหายร้อยละ 5 และร้อยละ10 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 LD EPR และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 20 LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อข้อมูลสูญหายร้อยละ 5 และร้อยละ 10 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 และร้อยละ 20 LD และ IWJR มีประสิทธิภาพสูงที่สุด

2. ไม่พบความแตกต่างของอำนาจการทดสอบ ของวิธีการประมาณข้อมูลสูญหายทั้ง 4 วิธี

3. พบปฏิสัมพันธ์สามทาง ในการประมาณค่าเฉลี่ยประชากร 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย และ 2) ระดับความสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย ในการประมาณความแปรปรวนประชากร 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย 2) ระดับความสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย และ 3) ขนาดตัวอย่าง ร้อยละข้อมูลสูญหาย วิธีการประมาณข้อมูลสูญหาย ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย 2) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและวิธีการประมาณข้อมูลสูญหาย และ3) ระดับสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย

4. ในการประมาณค่าเฉลี่ยประชากร ความแปรปรวนประชากรและสัมประสิทธิ์สหสัมพันธ์ประชากร IWJRมีความแกร่งต่อขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย ซึ่งสอดคล้องกันทั้งข้อมูลจากการจำลองและข้อมูลจริง

คำสำคัญ : ข้อมูลสูญหาย ข้อมูลสูญหายแบบสุ่มอย่างสมบูรณ์ การสุ่มตัวอย่างแบบง่าย

Abstract

The purpose of this study were first to develop the iterated weighted Jacknife method and regression (IWJR) for missing data estimation, and secondly to compare its efficiency of estimation population mean, population variance, and population corretlion and statistical power of a test under missing complete at random (MCAR) and simple random sampling with another four well-defined methods, namely; Listwise deletion (LD), Mean imputation (MI) and Regression imputation (RI). By using simulation and secondary data, the comparisons were made with the following conditions: i) three sample size (100, 200 and 500) ii) three level of correlation of variables (low ρ = .3 moderate ρ = .5 and high ρ = .7), and iii) four level of percentage of missing data (5%, 10%, 15% and 20%). Moreover, with a simulation data, a study of interaction among sample size, correlation of variables, number of missing data and missing data methods were also performed, at significant level .05.

Hence, the studies are summaried by the following items:

1. Used simulation data 1) when classified according to sample size. Under the population mean estimation i) 100 and 200 samples RI and IWJR method performed most effectively, and ii) 500 samples LD and IWJR performed most effectively. Under the population variance and population correlation estimation i) 100 samples IWJR method performed most effectively, and ii) 200 and

500 samples LD and IWJR performed most effectively. 2) When classified according to correlation of variables. Under the population mean estimation i) low correlation MI RI and IWJR method performed most effectively, and ii) moderate and high correlation, RI and IWJR performed most effectively. Under the population variance estimation low, moderate and high correlation LD and IWJR performed most effectively. Under the population correlation estimation i) low correlation IWJR method performed most effectively, ii) moderate correlation, LD and IWJR performed most effectively, and iii) high correlation, MI RI and IWJR performed most effectively. 3) When classified according to percentage of missing data. Under the population mean estimation i) 5% of missing data, IWJR performed most effectively, ii) 10% of missing data, RI and IWJR performed most effectively, and iii) 15% and 20% of missing data, LD MI RI and IWJR performed most effectively. Under the population variance estimation i) 5% and 10% of missing data, IWJR performed most effectively, ii) 15% of missing data, LD and IWJR performed most effectively, and iii) 20% of missing data, LD and IWJR performed most effectively. Under the population correlation estimation i) 5% and 10% of missing data, IWJR performed most effectively, and ii) 15% and 20% of missing data, LD and IWJR performed most effectively.

2. There was non-significant in the statistical power testing of all five methods of missing data.

3. Three-way interactions were found: under the population mean estimation i) sample size, correlation of variable, and percentage of missing data and ii) correlation of variable percentage of missing data, and methods of missing. Under the population variance estimation i) sample size, correlation of variables, and percentage of missing data ii) correlation of variables, percentage of missing data, and methods of missing data and iii) sample size, percentage of missing data, methods missing data. Under the population correlation estimation i) sample size, correlation of variables, and percentage of missing data ii) sample size, correlation of variable, and methods of missing and iii) correlation of variable, percentage of missing data, and methods of missing data.

4. Under the different estimations of population mean, variance, and correlation, the IWJR method demonstrated correspondingly well in both simulation and primary data, and presented robustness within the following conditions: i) the sample size, ii) the correlation of variable, and iii) the percentage of missing data.

Keyword: missing data, MCAR, simple random sampling

ภูมิหลัง

ข้อมูลสูญหายเป็นปัญหาที่พบโดยทั่วไปในการวิจัยเชิงปริมาณ (Heeringa, 2000) เป็นเรื่องที่ต้องเกิดขึ้นที่นักวิจัยจะต้องเผชิญ แม้ว่าจะได้ควบคุมการสำรวจหรือการทดลองไว้แล้วเป็นอย่างดีก็ตาม (Huisman. 1998: 271-278) นักวิจัยได้ให้ความสำคัญกับปัญหาข้อมูลสูญหายและให้ความสำคัญเพิ่มขึ้นเรื่อย ๆ (Adam. 2001) ในการวิเคราะห์ข้อมูลตัวแปรพหุ ถ้าตัวแปรแต่ละตัวมีข้อมูลหายไปโดยสุ่มเพียงร้อยละ 10 จะมีผลให้ต้องตัดหน่วยวิเคราะห์ทิ้งถึงร้อยละ 59 (Kim and Curry. 1977 อ้างถึงใน Roth. 1995: 1003-1023) การวิเคราะห์ข้อมูลจากเฉพาะข้อมูลที่เหลืออยู่ภายหลังจากที่ได้ตัดทิ้งค่าสังเกตที่ไม่สมบูรณไปแล้ว ผลการวิเคราะห์จะเอนเอียงและไม่ถูกต้อง (Wang. 2000) การประมาณค่าข้อมูลสูญหายเพื่อทดแทนข้อมูลที่สูญหายไปนั้นทำให้ประสิทธิภาพของการประมาณค่าและการสรุปผลการวิจัยสูงขึ้นอย่างน่าประหลาดใจ (Raymond, M.R. 1986: 395-420) การประมาณค่าข้อมูลสูญหายเพื่อทดแทนข้อมูลที่สูญหายไปนั้น ให้ผลแตกต่างจากการละเลยไม่สนใจกับข้อมูลที่สูญหายไปอย่างเห็นได้ชัด ซึ่งส่งผลโดยตรงกับคุณภาพของงานวิจัยทางการศึกษา ดังนั้นนักวิจัยทางการศึกษาควรให้ความสำคัญต่อการประมาณค่าข้อมูลสูญหายเพื่อให้ได้งานวิจัยทางการศึกษาที่มีคุณภาพมากขึ้น (Peng, C.-Y. J., et al. 2006: 31-78) ในทฤษฎีการตอบสนองข้อสอบ เมื่อร้อยละของข้อมูลสูญหายเพิ่มขึ้น การแสดงสาเหตุที่ผิดพลาดในการวิเคราะห์เชิงสาเหตุและยังส่งผลให้แบบจำลองจากการวัดไม่สอดคล้องกับแบบจำลองตามทฤษฎี (Zhang, B. and Walker, C.M. 2008: 466-479) ผลกระทบของข้อมูลสูญหายที่มีต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ ความเอนเอียงของการอธิบายความผันแปรของข้อสอบ ความคลาดเคลื่อนประเภทที่หนึ่งและความคลาดเคลื่อนประเภทที่สองได้รับผลกระทบโดยตรงและค่อนข้างจะมีผลกระทบที่รุนแรง (Robitszsch, A. and Rupp, A. A. 2009: 18-34) และการประมาณพารามิเตอร์ของข้อสอบส่งผลกระทบในทางลบและเพิ่มมากขึ้น เมื่อปริมาณข้อมูลสูญหายเพิ่มขึ้น (Furlow, CF. et al. 2007: 388-403) การจัดการข้อมูลสูญหายที่นักวิจัยใช้กันอยู่ในปัจจุบัน ได้แก่ LD ข้อบกพร่องของวิธีการนี้ คือ อำนาจการทดสอบลดลง ค่าประมาณค่าความแปรปรวนต่ำกว่าความเป็นจริง (Underestimate) (Little and Rubin. 2002) MI ข้อบกพร่องของวิธีการนี้ คือ ค่าประมาณค่าความแปรปรวนต่ำกว่าความเป็นจริง (Rovine and Delaney. 1990: 35–79; Landerman, Land & Pieper. 1997: 3–33; Brockmeier. 1998: 20-39) และ RI ข้อบกพร่องของวิธีการนี้ คือ ค่าประมาณค่าความแปรปรวนต่ำกว่าความเป็นจริง ทำให้เกิดภาวะร่วมเส้นตรงหลายตัวแปร (Multicolinearity) (Little and Schenker. 1995) นอกจากนี้วิธีการประมาณข้อมูลสูญหายด้วยค่าเพียงค่าเดียวสำหรับทุก ๆ ค่าที่สูญหายวิธีการเช่นนี้ทำให้ค่าความแปรปรวนที่ประมาณได้ต่ำกว่าความเป็นจริง

ความมุ่งหมายของการวิจัย

1. เพื่อพัฒนาวิธีการประมาณข้อมูลสูญหายโดยการถ่วงน้ำหนักแบบวนซ้ำด้วยวิธีของแจ็คไนฟ์และการวิเคราะห์การถดถอย

2. เพื่อเปรียบเทียบประสิทธิภาพ การประมาณค่าเฉลี่ยประชากร การประมาณค่าความแปรปรวนประชากร การประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร อำนาจการทดสอบในการทดสอบสหสัมพันธ์และการวิเคราะห์ความแปรปรวน ของ IWJR LD MI และ RI จากขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหายที่แตกต่างกัน โดยใช้ข้อมูลจากการจำลองและข้อมูลจริง

3. เพื่อศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหายในการประมาณค่าเฉลี่ยประชากร ความแปรปรวนประชากรและสัมประสิทธิ์สหสัมพันธ์ประชากร โดยใช้ข้อมูลจากการจำลอง

ขอบเขตของการวิจัย

ขอบเขตของการวิจัยในครั้งนี้มีดังต่อไปนี้

1. ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลที่ได้จากการจำลองสถานการณ์ โดยใช้เทคนิค มอนติคาร์โลซิมูเลชั่น (Monte carlo simulation) และข้อมูลจริงใช้ข้อมูลทุติยภูมิ (Secondary data) จากฐานข้อมูลของสำนักงานกรรมการศึกษาขั้นพื้นฐาน เป็นข้อมูลผลการเรียนเฉลี่ยรวม (GPAX) และผลการเรียนเฉลี่ยกลุ่มสาระ (GPA) จำนวน 2,381 โรงเรียน

2. การเปรียบเทียบประสิทธิภาพของวิธีการประมาณข้อมูลสูญหายพิจารณาจากค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean square error: MSE) (Hank John E., Reitsch Arthur G. and Wichern Dean W. 2001: 75) ดังนี้ $MSE = \sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2 \Big/ n$

3. ตัวแปรที่ศึกษามีดังต่อไปนี้

3.1 ตัวแปรอิสระ 1) วิธีการประมาณข้อมูลสูญหาย 4 วิธี คือ LD MI RI และ IWJR 2) ขนาดตัวอย่าง 100 200 และ 500 (Timm. 1970: 417-437; Beale and Little. 1975: 129-145; Gleason and Staelin. 1975: 229-252; Chan, Gilman and Dunn. 1976: 842-844; Little. 1976: 593-604; Raymond and Roberts. 1987: 13-26; Viragoontavan, 2000; Chaimongkol and Suwattee, 2004; เชาว์ อินใย. 2552) 3) ความสัมพันธ์ระหว่างตัวแปร 3 ระดับ คือ ต่ำ ปานกลางและสูง (Frane. 1976: 409–415; Little and Rubin. 2002; Hegamin-Younger and Forsyth. 1988: 197-210; Brokckmeier, Kromrey and Hines. 1988: 20-39; Viragoontavan. 2000; Chaimongkol and Suwattee. 2004; เชาว์ อินใย. 2552) และ 4) ข้อมูลสูญหาย ร้อยละ 5 10 15 และ 20 (Roth. 1994; Viragoontavan. 2000; เชาว์ อินใย. 2552)

3.2 ตัวแปรตาม ได้แก่ ประสิทธิภาพของวิธีการประมาณข้อมูลสูญหาย อำนาจการทดสอบจากการใช้การทดสอบสหสัมพันธ์และการวิเคราะห์ความแปรปรวนและปฏิสัมพันธ์ระหว่าง วิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปร และร้อยละข้อมูลสูญหาย

วิธีดำเนินการวิจัย

IWJR มีขั้นตอนดังนี้

ขั้นตอนที่ 1 จากข้อมูลที่สมบูรณ์ $y_1, y_2, \cdots, y_r$ นำมาหาค่าเฉลี่ยของข้อมูลโดยใช้วิธีสุ่มซ้ำโดยวิธีของแจ็คไนฟ์ (Quenouille. 1956: 353-360) ค่าประมาณจะถูกกำหนดโดย $\overline{y}_j = \sum_{i=1}^{r}\overline{y}_i \Big/ r \;;\; j = r+1, r+2, \cdots, n$ (Yu Chong Ho. 2003; Suat SAHINLER and Derviz TOPUZ. 2007: 188-199; ปรีชา วิจิตรธรรมรส. 2542 : 13-21)

ขั้นตอนที่ 2 จากข้อมูลที่สมบูรณ์ นำมาสร้างสมการถดถอย กำหนดให้ $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j$ สำหรับ $r+1 \le j \le n$ (Draper, Norman R., Smith Harry. 1998: 22-24)

ขั้นตอนที่ 3 นำค่า $\overline{y}_j$ จากขั้นตอนที่ 1 และค่า $\hat{y}_j$ จากขั้นตอนที่ 2 มาทำการถ่วงน้ำหนัก จากสมการ $\hat{y}_j^* = w_j\overline{y}_j + (1-w_j)\hat{y}_j$ เมื่อ

$$w_j = \hat{\sigma}^2_{Reg.}\left[\frac{1}{r} + (x_j - \overline{x})^2 \Big/ \sum_{j=1}^{r}(x_j - \overline{x})^2\right] \Big/ \frac{\hat{\sigma}^2_{jackknife}}{r} + \hat{\sigma}^2_{reg}\left[\frac{1}{r} + (x_j - \overline{x})^2 \Big/ \sum_{i=1}^{r}(x_i - \overline{x})^2\right]$$

สุ่มข้อมูลสูญหายขึ้นมา 1 ค่า นำค่า $\hat{y}_j^*$ ทดแทนข้อมูลสูญหาย นำค่าทดแทนข้อมูลสูญหายที่ได้จากขั้นตอนที่ 3 รวมเข้ากับข้อมูลที่สมบูรณ์ ทำซ้ำขั้นตอนที่ 1 ถึง 3 จนครบทุกค่าของข้อมูลสูญหาย

การตรวจสอบประสิทธิภาพ มีรายละเอียดดังนี้

1. คำนวณค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ของประชากร
2. สุ่มตัวอย่างจากประชากร ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย โดยมีขนาดตัวอย่างตามที่กำหนด
3. สร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์ โดยมีร้อยละข้อมูลสูญหายตามที่กำหนด
4. ประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายทั้ง 4 วิธี
5. คำนวณค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์จากตัวอย่างของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี
6. คำนวณค่าความแตกต่างของค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์จากตัวอย่างของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี กับค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ของประชากร ทำซ้ำ 2) ถึง 6) จำนวน 1,000 ครั้ง (Chaimongkol & Suwattee, 2004; เชาว์ อินใย. 2547; เชาว์ อินใย. 2552) คำนวณค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสองของค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี

การหาอำนาจการทดสอบจากการทดสอบสัมประสิทธิ์สหสัมพันธ์ มีขั้นตอนการดำเนินการดังนี้

1. สุ่มตัวอย่างจากประชากร ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย โดยมีขนาดตัวอย่างตามที่กำหนด
2. สร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์ โดยมีร้อยละข้อมูลสูญหายตามที่กำหนด
3. ประมาณข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายทั้ง 4 วิธี
4. ทดสอบสมมุติฐานสัมประสิทธิ์สหสัมพันธ์ของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี โดยมีสูตรดังนี้ (Richard J. R. and Marx, Morris L. 1986: 437) $t = r\sqrt{n-2/1-r^2}$, $df = n-2$

$$r = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \Bigg/ \sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}$$

ทำซ้ำ 1) ถึง 4) จำนวน 1,000 ครั้ง หาอำนาจการทดสอบของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี

การหาอำนาจการทดสอบจากการทดสอบความแตกต่างของค่าเฉลี่ย มีขั้นตอนการดำเนินการดังนี้

1. สร้างข้อมูลประชากร 3 กลุ่ม ที่มีระดับความสัมพันธ์ระหว่างตัวแปรตามที่กำหนด โดยให้มีค่าเฉลี่ยแตกต่างอย่างมีนัยสำคัญทางสถิติ ที่ระดับนัยสำคัญ .05 และให้ความแปรปรวนทั้ง 3 กลุ่มเท่ากัน
2. สุ่มตัวอย่างจากประชากรทั้ง 3 กลุ่ม ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย โดยมีขนาดตัวอย่างตามที่กำหนด
3. สร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์ โดยให้มีร้อยละของข้อมูลสูญหายตามที่กำหนดทั้ง 3 กลุ่ม
4. ประมาณข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายทั้ง 4 วิธี ทั้ง 3 กลุ่ม 4) ทดสอบสมมุติฐานความแตกต่างระหว่างค่าเฉลี่ยของ 3 กลุ่ม ของวิธีการประมาณข้อมูลสูญหายแต่ละวิธี โดยใช้การวิเคราะห์ความแปรปรวนโดยใช้สถิติ F และทดสอบความแตกต่างด้วยวิธีการของตูกี้

ตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย ในการประมาณค่าเฉลี่ยประชากร ความแปรปรวนประชากรและสัมประสิทธิ์สหสัมพันธ์ประชากร โดยใช้วิธีการวิเคราะห์ความแปรปรวน

สรุปผล

1. ผลการวิจัยพบว่า IWJR มีประสิทธิภาพสูงที่สุด ที่ระดับนัยสำคัญ .05 ภายใต้เงื่อนไขดังต่อไปนี้ ในการประมาณค่าเฉลี่ยประชากร 1) ขนาดตัวอย่าง 100 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง และข้อมูลสูญหายร้อยละ 5 2) ขนาดตัวอย่าง 200 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำและปานกลาง ข้อมูลสูญหายร้อยละ 5 และ 10 และ 3) ขนาดตัวอย่าง 500 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำและปานกลาง ข้อมูลสูญหายร้อยละ 5 ในการประมาณค่าความแปรปรวนประชากร 1) ขนาดตัวอย่าง 100 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำและปานกลาง ข้อมูลสูญหายร้อยละ 5 2) ขนาดตัวอย่าง 200 และ 500 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง ข้อมูลสูญหายร้อยละ 5 ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร ขนาดตัวอย่าง 100 200 และ 500 ระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง ข้อมูลสูญหายร้อยละ 5

2. เมื่อจำแนกตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ .05 ในการประมาณค่าเฉลี่ยประชากร เมื่อขนาดตัวอย่าง 100 และ 200 RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อขนาดตัวอย่าง 500 LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากรและค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อขนาดตัวอย่าง 100 IWJR มีประสิทธิภาพที่สูงที่สุด เมื่อขนาดตัวอย่าง 200 และ 500 LD และ IWJR มีประสิทธิภาพสูงที่สุด

3. เมื่อจำแนกตามระดับความสัมพันธ์ระหว่างตัวแปร ที่ระดับนัยสำคัญ .05 ในการประมาณค่าเฉลี่ยประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ MI RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรปานกลางและสูง RI และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรปานกลาง LD และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรสูง MI RI และ IWJRมีประสิทธิภาพสูงที่สุด

4. เมื่อจำแนกตามร้อยละข้อมูลสูญหาย ที่ระดับนัยสำคัญ .05 ในการประมาณค่าเฉลี่ยประชากร เมื่อข้อมูลสูญหายร้อยละ 5 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 10 RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 และร้อยละ 20 LD MI RI และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนประชากร เมื่อข้อมูลสูญหายร้อยละ 5 และร้อยละ10 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 LD และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 20 LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร เมื่อข้อมูลสูญหายร้อยละ 5 และร้อยละ 10 IWJR มีประสิทธิภาพสูงที่สุด เมื่อข้อมูลสูญหายร้อยละ 15 และร้อยละ 20 LD และ IWJR มีประสิทธิภาพสูงที่สุด

5. ไม่พบความแตกต่างของอำนาจการทดสอบ ที่ระดับนัยสำคัญ .05 ของวิธีการประมาณข้อมูลสูญหายทั้ง 4 วิธี

6. พบปฏิสัมพันธ์สามทาง ที่ระดับนัยสำคัญ .05 ในการประมาณค่าเฉลี่ยประชากร ระหว่าง 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย และ 2) ระดับความสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย ในการประมาณค่าความแปรปรวนประชากร ระหว่าง 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย 2) ระดับความสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย และ 3) ขนาดตัวอย่าง ร้อยละข้อมูลสูญหาย วิธีการประมาณข้อมูลสูญหาย ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์ประชากร ระหว่าง 1) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละ

ข้อมูลสูญหาย 2) ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและวิธีการประมาณข้อมูลสูญหาย และ3) ระดับสัมพันธ์ระหว่างตัวแปร ร้อยละข้อมูลสูญหายและวิธีการประมาณข้อมูลสูญหาย

7. ในการประมาณค่าเฉลี่ย ความแปรปรวนประชากรและสัมประสิทธิ์สหสัมพันธ์ประชากร IWJR มีความแกร่งต่อขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย ซึ่งสอดคล้องกันทั้งข้อมูลจากการจำลองและข้อมูลจริง

ตาราง 1 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย การประมาณค่าเฉลี่ย ค่าความแปรปรวนประชากรและค่าสัมประสิทธิ์สหสัมพันธ์ประชากร จำแนกตามขนาดตัวอย่าง โดยใช้ข้อมูลจากการจำลอง

การประมาณ	ขนาดตัวอย่าง	วิธีการประมาณข้อมูลสูญหาย			
		LD	MI	RI	IWJR
ค่าเฉลี่ยประชากร	100	.00068403*	.00068403*	.00065976	**.00063048**
	200	.00033356*	.00033356*	.00032294	**.00030374**
	500	.00013805	.00013480*	.00013322*	**.00012276**
ความแปรปรวนประชากร	100	.00008149*	.00012973*	.00010601*	**.00007650**
	200	.00003946	.00009749*	.00007044*	**.00003809**
	500	**.00001716**	.00007913*	.00005073*	.00001720
สัมประสิทธิ์สหสัมพันธ์ประชากร	100	.00645329*	.00762241*	.00722540*	**.00582988**
	200	.00306395	.00433727*	.00378934*	**.00277454**
	500	**.00132508**	.00270702*	.00198707*	.00132970

ตาราง 2 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย การประมาณค่าเฉลี่ย ค่าความแปรปรวนประชากรและค่าสัมประสิทธิ์สหสัมพันธ์ประชากร จำแนกตามระดับความสัมพันธ์ โดยใช้ข้อมูลจากการจำลอง

การประมาณ	ระดับความสัมพันธ์	วิธีการประมาณข้อมูลสูญหาย			
		LD	MI	RI	IWJR
ค่าเฉลี่ยประชากร	ต่ำ	.00039050*	.00038725	.00038690	**.00035877**
	ปานกลาง	.00038274*	.00038274*	.00036980	**.00034740**
	สูง	.00038240*	.00038240*	.00035921	**.00035081**
ความแปรปรวนประชากร	ต่ำ	.00004832	.00010399*	.00009354*	**.00004525**
	ปานกลาง	.00004245	.00009717*	.00007282*	**.00003990**
	สูง	.00004734	.00010520*	.00006082*	**.00004664**
สัมประสิทธิ์สหสัมพันธ์ประชากร	ต่ำ	.00543328*	.00529314*	.00639580*	**.00472743**
	ปานกลาง	.00375524	.00485996*	.00452457*	**.00358388**
	สูง	.00165379*	.00451360	.00208144	**.00162280**

ตาราง 3 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย การประมาณค่าเฉลี่ย ค่าความแปรปรวนประชากรและค่าสัมประสิทธิ์สหสัมพันธ์ประชากร จำแนกตามร้อยละข้อมูลสูญหาย โดยใช้ข้อมูลจากการจำลอง

การประมาณ	ร้อยละข้อมูลสูญหาย	วิธีการประมาณข้อมูลสูญหาย			
		LD	MI	RI	IWJR
ค่าเฉลี่ยประชากร	5	.00034240*	.00034240*	.00033847*	**.00027604**
	10	.00036897*	.00036897*	.00036021	**.00033437**
	15	.00038958	.00038525	**.00037199**	.00037392
	20	.00043991	.00043991	**.00041722**	.00042498
ความแปรปรวนประชากร	5	.00004186*	.00004601*	.00004378*	**.00003406**
	10	.00004388*	.00007144*	.00005808*	**.00003863**
	15	**.00004618**	.00011371*	.00008221*	.00004790
	20	**.00005223**	.00017732*	.00011884*	.00005512
สัมประสิทธิ์สหสัมพันธ์ประชากร	5	.00320011*	.00329692*	.00332795*	**.00256962**
	10	.00351932*	.00420239*	.00386478*	**.00293524**
	15	.00367286	.00510648*	.00457816*	**.00365757**
	20	**.00406413**	.00694981*	.00556484*	.00408304

ตัวเลขตัวหนา หมายถึง มีประสิทธิภาพสูงสุด * หมายถึง ค่าเฉลี่ยประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุดที่ระดับนัยสำคัญ .05

ข้อเสนอแนะ

1. ศึกษาภายใต้ประเภทการสูญหายแบบสุ่ม (MAR) และแบบไม่สุ่ม (NMAR)

2. ศึกษาวิธีการประมาณค่าข้อมูลสูญหายที่นักวิจัยพัฒนาขึ้น กับวิธีการประมาณค่าข้อมูลสูญหายอื่น ๆ เช่น วิธีการประมาณข้อมูลสูญหายแบบฮ็อทเดสก์ วิธีการประมาณข้อมูลสูญหายแบบฮ็อทเดสก์เชิงสุ่ม วิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์-ไนจ์บอร์ ฮ็อทเดสก์ วิธีการประมาณข้อมูลสูญหายด้วยวิธีอีเอ็ม และวิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ

3. ศึกษาโดยใช้ข้อมูลที่มีตัวแปรอิสระมากกว่าหนึ่งตัว

4. ใช้ทริมด์มีนส์ (Trimmed Means) เข้าร่วมในการพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย

5. ใช้สมการถดถอยที่มีวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีอื่นๆ เข้าร่วมในการพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย

บรรณานุกรม

เชาว์ อินใย. (2547). *การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีและการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวท์:เทคนิคมอนติคาร์โล*. วิทยานิพนธ์ปริญญาเอก. พิษณุโลก : มหาวิทยาลัยนเรศวร.

เชาว์ อินใย. (2552). *การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอาร์และการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวท์:เทคนิคมอนติคาร์โล*. เลย : มหาวิทยาลัยราชภัฏเลย.

ปรีชา วิจิตรธรรมรส. (2542). ตัวประมาณแจ็คไนฟ์. *วารสารพัฒนบริหารศาสตร์*, ปีที่ 39(ฉบับที่ 3), กรุงเทพฯ : สถาบันบัณฑิตพัฒนบริหารศาสตร์. ก.ค.-ก.ย. 2542 : หน้า 13-21.

Adam, Carlson. (2001). Data Mining: Finding Nuggets of Knowledge in Mountains of Data. *Northwest Science & Technology*, Autumn, 24-25.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, 37, 129-145, B.

Brockmeier, L. L., Kromrey, J. D. & Hines, C. V. (1998). Systematically missing data and multiple regression analysis: An empirical comparison of deletion and imputation techniques. *Multiple Linear Regression Viewpoints*, 25, 20-39.

Chaimongkol , W. (2004). *Three composite imputation methods for item nonresponse estimation in sample surveys*. Doctor's Thesis. Bangkok :National Institute of Development Administration.

Chan, L. S., Gilman, J. A., & Dunn, O. J. (1976). Alternative approaches to missing values in discriminant analysis. *Journal of the American Statistical Association*, 71, 842-844.

Draper, Norman R., Smith Harry. (1998). *Applied Regression Analysis*. 3rd ed. John Willey & Sons, Inc. NY.

Frane, J.W. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*, 41, 409–415.

Furlow CF, et al. (2007). A Monte Carlo study of the impact of missing data and differential item functioning on theta estimates from two polytomous Rasch family models. *Journal of Applied Measurement*, 8(4), 388-403.

Gleason, T. C. & Staelin, R. (1975). A proposal for handling missing data. *Psvchometrika*, 40, 229-252.

Hank John E., Reitsch Arthur G. and Wichern Dean W. (2001). *Business Forecasting*, 7th ed. New Jersey. Prentice Hall.

Hegamin-Younger, C. & Forsyth, R. (1998). A comparison of four imputation procedures in a two-variable prediction system. *Educational and Psychological Measurement*, 58(2), 197-210.

Huisman, M. (1998). Item Nonrespons : Occurrence cause, and Imputation of Missing Answers to Test Item. *DSWO Press, Lieden University, The Netherlands*.

Landerman LR, Land KC, Pieper CF. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods and Research*, 26(1), 3–33.

Little, R. J. A. & Schenker, N. (1995). Missing data. *In G. Arminger, C. C. Clogg, & M.E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences*. New York.

Little, R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrika*, 63, 593-604.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons.

Peng, C.-Y. J., et al. (2006). Advances in missing data methods and implications for educational research In Sawilowsky, S. (eds). *Real data analysis*. Greenwich, CT., Information Age Publishing Inc. 31-78.

Quenonuille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*. 43, 353-360.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, 47, 13-26.

Raymond, M.R. (1986). Missing Data in Evaluation Research. *Eval Health Prof*. 9(4), 395-420.

Richard J. R. and Marx, Morris L. (1986). *An Introduction to Mathematical Statistics and Its Applications*, New Jersey. Prentice-Hall.

Robitszsch, A. and Rupp, A. A. (2009). Impact of Missing Data on the Detection of Differential Item Functioning: The Case of Mantel-Haenszel and Logistic Regression Analysis. *Educational and Psychological Measurement*, 69(1), 18-34.

Roth, P.L. (1994). Missing Data: A Conceptual Review for Applied Psychology. *Journal of Personal Psychology*, 47, 537-560.

Rovine, M. J., & Delaney, M. (1990). Missing data estimation in developmental research. *In A. Von Eye (Ed.), Statistical methods in longitudinal research, Stanford: Academic Press*, 1, 35–79.

Suat SAHINLER and Derviz TOPUZ. (2007). Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters. *Journal of Applied Qualitative Methods*, 2(2), Summer 2007, 188-199.

Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*. 35(4), 417-437.

Viragoontavan, S. (2000). *Comparing Six Missing Data Methods within the Discriminant Analysis Context: A Monte Carlo Study*. Doctor's Thesis. Ohio : The Ohio State University.

Wang, Betty Lu-Ti. (2000). *Imputation Methods for missing Data in Growth Curve Models*. Doctor's Thesis. California : University of Southern California. Dissertation Abstract International. < http://proquest.umi.com/pqdweb?did =728849541 &sid=2&Fmt= 2&clientId=73599&RQT=309&VName=PQD> October, 13 2009.

Yu, Chong Ho. (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, 8(19). http://pareonline.net/getvn.asp? v=8&n=19 October, 13 2009.

Zhang, B. and Walker, C.M. (2008). Impact of Missing Data on Person Model Fit and Person Trait Estimation. *Applied Psychological Measurement*, 32(8), 466-479.