สุพจน์ นาคสวัสดิ์
54810018
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

การจัดการปัญหาข้อมูลสูญหายในงานวิจัยด้วยโปรแกรมลิสเรล (LISREL)

(Method of handling Missing Data by LISREL)

ข้อมูลสูญหาย (missing data) คือ ค่าสังเกตที่ต้องการทราบค่าแต่ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะสามารถทราบค่าได้หากวิธีการที่ใช้การเก็บรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีขึ้นหรือมีความเหมาะสมมากขึ้น เป็นกรณีที่พบได้บ่อยในงานวิจัยทุกสาขา และนักวิจัยจำเป็นต้องพิจารณาถึงแนวทางที่เหมาะสมสำหรับใช้จัดการกับข้อมูลสูญหายในทุก ๆ ครั้งที่พบกับปัญหานี้ ซึ่งวิธีการที่ใช้สำหรับจัดการกับข้อมูลสูญหายมีทางเลือกให้พิจารณาค่อนข้างหลากหลาย หากเลือกใช้วิธีจัดการกับข้อมูลสูญหายที่ไม่เหมาะสมย่อมส่งผลทำให้เกิดการบิดเบือนผลการวิเคราะห์ อย่างไรก็ตามจากการศึกษาของ Wood และคณะ (2004) ที่ได้ทำการศึกษาจากผลงานวิจัยที่ได้รับการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet และ N Engl J Med จำนวนน 71 ชิ้น พบว่า มีงานวิจัยถึง 89% ที่มีปัญหาเรื่องข้อมูลสูญหาย และมีเพียง 21% เท่านั้นที่มีการจัดการกับปัญหาข้อมูลที่ไม่สมบูรณ์ จากผลการศึกษานี้แสดงให้เห็นว่า การจัดการกับปัญหาข้อมูลสูญหายยังคงถูกละเลยกันอย่างเป็นปกติ (ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี, 2006)

- **เหตุผลหลักของการเกิดข้อมูลสูญหาย**

  เกิดจากหลายกรณี โดยเหตุผลพื้นฐานมักเป็นผลจาก

  1. หน่วยตัวอย่างปฏิเสธการตอบคำถามในบางคำถามของแบบฟอร์มที่ใช้รวบรวมข้อมูล ซึ่งส่วนใหญ่มักเป็นคำถามที่กระทบต่อความรู้สึกได้ง่าย
  2. หน่วยตัวอย่างไม่ทราบคำตอบ ทั้งนี้อาจเป็นผลมาจากปัญหาในเรื่องของความจำ
  3. คำถามที่ใช้ไม่ครอบคลุมทุกกรณีจึงทำให้เกิดข้อมูลสูญหาย
  4. ความบกพร่องของแบบสอบถาม เช่น จำนวนข้อคำถามเยอะเกินไป แบบสอบถามที่พิมพ์หน้าหลัง
  5. ความผิดพลาดจากหน่วยตัวอย่าง เช่น ลืม กรอกข้อมูลผิดพลาด เป็นต้น
  6. การนำข้อมูลเข้าสู่ระบบประมวลผลทางคอมพิวเตอร์ ซึ่งอาจเป็นผลจากความผิดพลาดของระบบฐานข้อมูล หรือผู้บันทึกข้อมูล เป็นต้น

- **การตรวจเช็คข้อมูลสูญหาย (to detect missing data)**

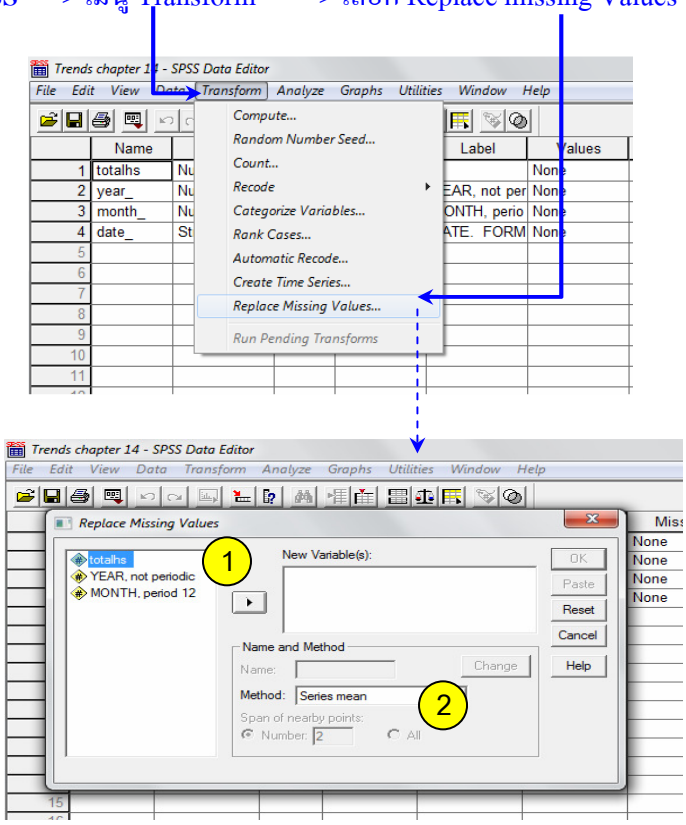
การตรวจสอบข้อมูลสูญหายภายหลังกรอกข้อมูล เช่น การหา Min และ Max แล้วจัดเรียง(Sort) ข้อมูล เมื่อพบข้อมูลหรือค่าที่ผิดก็ตรวจสอบกับแบบสอบถามชุดดังกล่าวว่าเกิดจากการกรอกข้อมูลผิดหรือผู้ตอบแบบสอบถามกรอกข้อมูลไม่ถูกต้อง สุดท้ายก็ให้แก้ไขกรณีที่กรอกผิด ในกรณีที่กรอกถูกต้องแต่ผู้ตอบแบบสอบถามกรอกผิด เช่น เลข 4 แต่ผู้ตอบแบบสอบถามกรอกเป็น 44 แบบนี้ก็ถือว่าเป็น Missing data เช่นกัน

สำหรับวิธีการตรวจเช็คข้อมูลสูญหายสามารถทำได้หลายวิธีด้วยกัน โดยในผลงานตีพิมพ์ของ O'Rourke (2000) ได้อธิบายวิธีการสำหรับตรวจเช็คข้อมูลสูญหายไว้อย่างละเอียดมาก ในที่นี้จะขอกล่าวถึงอย่างย่อ ๆ เกี่ยวกับวิธีการต่าง ๆ ที่สามารถใช้ในการตรวจเช็คข้อมูล ได้แก่ การตรวจเช็คด้วยสายตา (visual scanning) การใช้โปรแกรมนำเข้าข้อมูล (data entry program) เช่น QPL หรือ SPSS ช่วยในการตรวจเช็ค การแจกแจงความถี่ของคำตอบในตัวแปรแต่ละตัว และการวิเคราะห์ตัวแปรคู่ (bivariate analysis) ใช้วิธีการสร้างตารางไขว้ (cross-tablation) ระหว่างตัวแปรทั้งคู่

กรณีข้อมูลสูญหายเกิดจากผู้ตอบไม่ตอบแบบสอบ ทั้งตั้งใจและไม่ตั้งใจ (ลืม) โดยปกตินักวิจัยก็อาจจะใช้วิธีตัดทิ้งทั้งหมด (Pairwise) หรือตัดเฉพาะข้อที่ไม่ตอบ ซึ่งจะทำจำนวนผู้ตอบแบบสอบถามไม่เท่ากัน (n) แต่การตัดสินใจตัดข้อมูลสูญหายหรือไม่มีหลายเงื่อนไขที่เกี่ยวข้อง เช่น วัตถุประสงค์การวิจัย ตัดแล้วทำให้จำนวนกลุ่มตัวอย่างต่ำกว่าเกณฑ์หรือไม่ (ดังนั้นนักวิจัยจึงมักวางแผนเก็บข้อมูลมากกว่าขนาดตัวอย่างขั้นต่ำที่กำหนดเพื่อป้องกันปัญหาดังกล่าว)

อย่างไรก็ตามในกรณีที่ผู้วิจัยไม่ใช้วิธีตัดข้อมูลที่สูญหาย แต่ใช้วิธีการทดแทนข้อมูลที่สูญหาย (Replace missing data) ก็สามารถทำได้หลายแนวทาง เช่น การแทนข้อมูลสูญหายด้วยค่าเฉลี่ย (Replacement missing data with mean) ด้วยโปรแกรม SPSS ดังภาพประกอบด้านล่าง

เปิดโปรแกรม SPSS ----> เมนู Transform ------> เลือก Replace missing Values



จากหน้าต่าง Replace Missing Values ของ SPSS-----> (1) เลือกตัวแปรที่จะแทนข้อมูลสูญหายในช่อง New Variable(s):-----> (2) เลือก Series mean ในช่อง Method -------> กด OK

อย่างไรก็ตามการแทนข้อมูลสูญหายด้วยค่าเฉลี่ย (Replace missing data with mean) มีทั้งข้อดีและข้อเสีย **ข้อดี**คือผู้วิจัยได้ตัวเลขหรือข้อมูลครบถ้วนและจำนวนตัวอย่างเท่ากัน (n) แต่**ข้อเสีย**คือการแทนค่าด้วยค่าเฉลี่ยนั้นจริงๆแล้วผู้ตอบแบบสอบถามจะตอบอย่างนั้นจริงๆหรือไม่ เช่น

ข้อคำถาม	ผู้ตอบแบบสอบถาม			
	นาย ก	นาย ข	นาย ค	นาย ง
1	4	1	4	4
2	5	?←Missing → ?		5
3	4	2	4	4
4	4	2	4	4

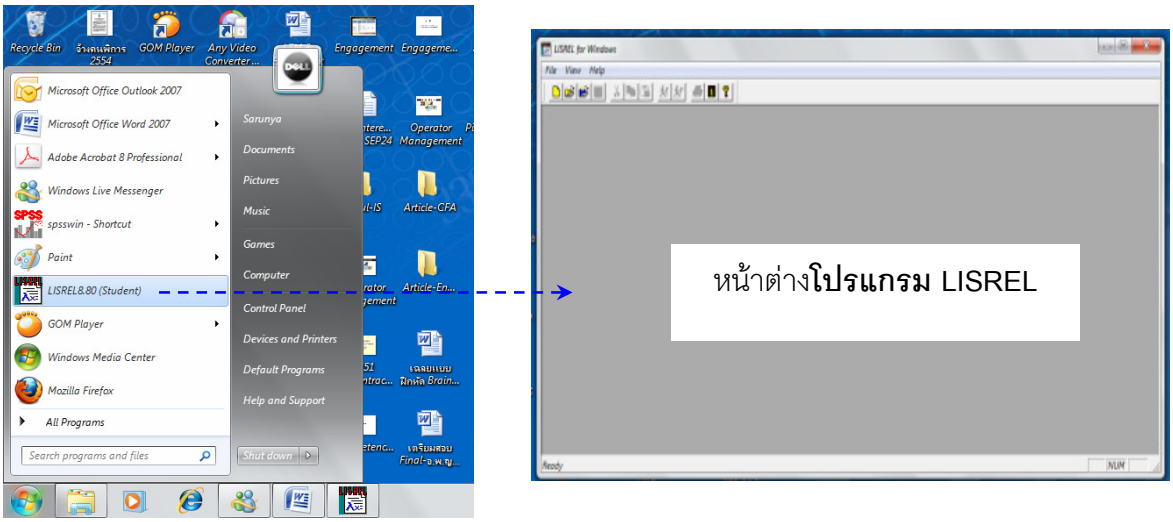
ดั้งนั้นค่าเฉลี่ยที่จะแทนให้กับนาย ข และนาย ค คือ 5 (เกิดจากค่าเฉลี่ยข้อคำถามข้อที่ 2 ที่คิดจากนาย ก และ นาย ง) ซึ่งดูแล้วการแทนตัวเลข 5 ข้อที่ 2 ให้กับนาย ข ดูแล้วไม่ค่อยสมเหตุสมผลเท่ากับนาย ค เพราะเห็นชัดว่าข้อมูลของนาย ข มีแนวโน้มจะเป็นไปในทิศทางต่ำๆ

จากข้อจำกัดดังกล่าวจึงเป็นที่มาของวิธีแทนข้อมูลสูญหายด้วยการจับคู่ (Matching) กับข้อมูลที่สมบูรณ์ของหน่วยตัวอย่างอื่นด้วย

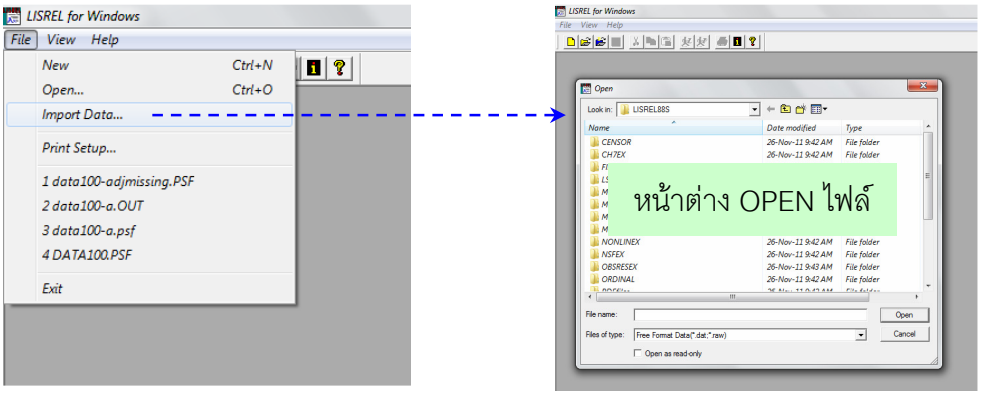
• วิธีทดแทนข้อมูลที่สูญหายด้วยวิธี Matching ด้วยโปรแกรมลิสเรล (LISREL)

ขั้นตอนการ Clean up ข้อมูลที่สูญหาย (Missing data) ประกอบด้วยขั้นตอนต่าง ๆ ดังนี้

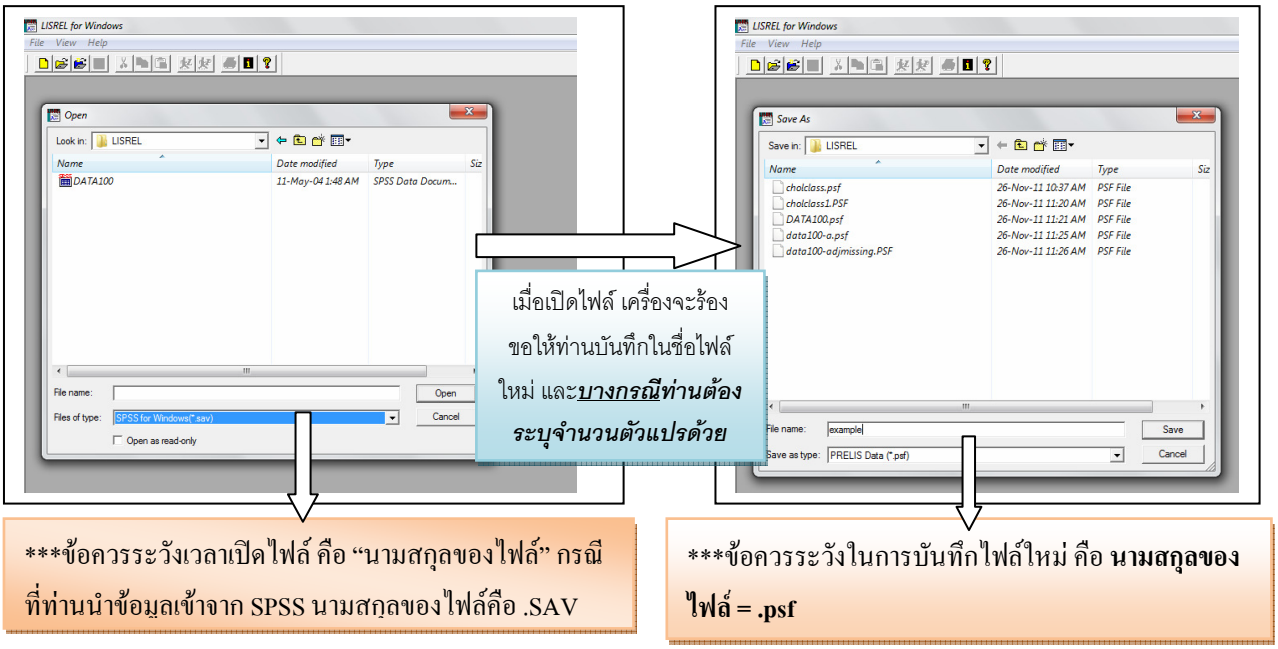
1. เปิดโปรแกรมลิสเรล (LISREL)



2. นำเข้าข้อมูล (Import data)



3.* เลือกไฟล์ เปิดไฟล์ และ*บันทึกไฟล์ใหม่

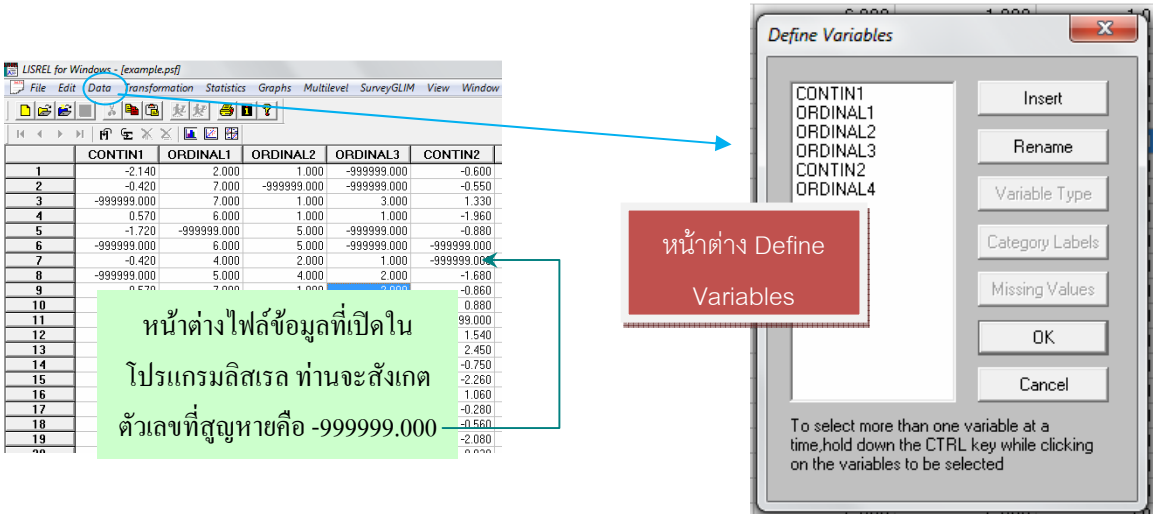


***ข้อควรระวังเวลาเปิดไฟล์ คือ “นามสกุลของไฟล์” กรณีที่ท่านนำข้อมูลเข้าจาก SPSS นามสกุลของไฟล์คือ .SAV

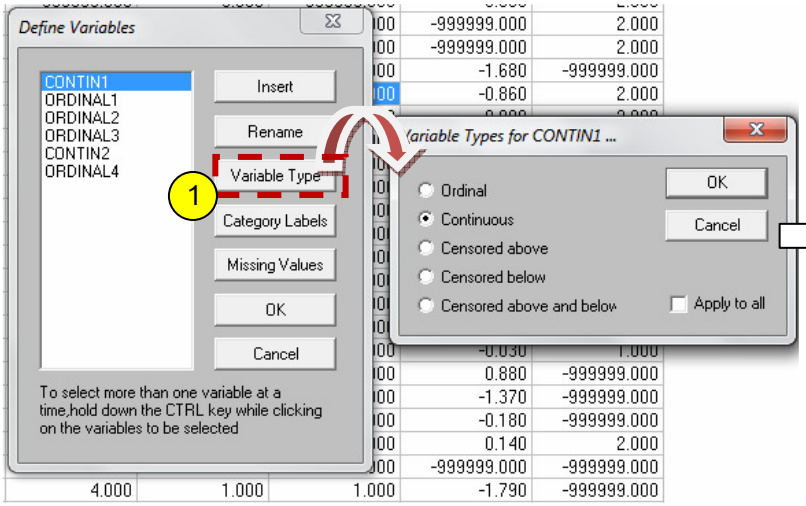
***ข้อควรระวังในการบันทึกไฟล์ใหม่ คือ **นามสกุลของไฟล์ = .psf**

4. **การกำหนดตัวแปร** (Define Variable)

เมื่อบันทึกไฟล์เรียบร้อยแล้ว หน้าต่างโปรแกรมลิสเรลก็จะแสดงข้อมูลไฟล์ที่ท่านต้องการ ซึ่งขั้นตอนต่อไปคือท่านต้องกำหนดตัวแปรเพื่อให้โปรแกรมรู้จักเสียก่อน

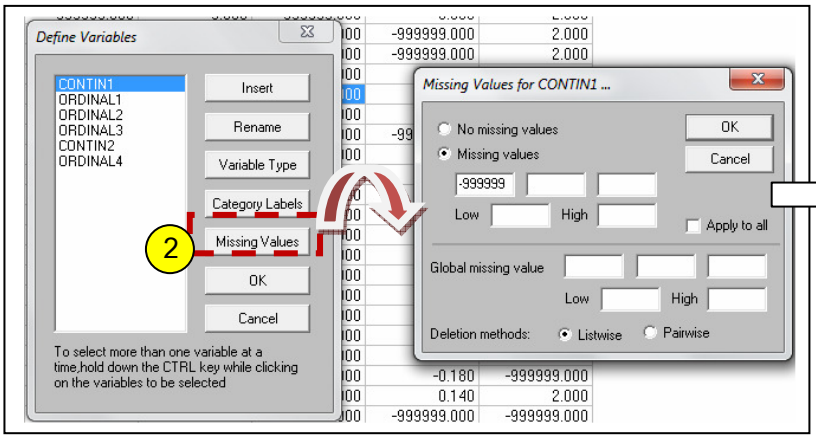


ในหน้าต่าง Define Variables ท่านจะต้อง (1) กำหนด Variable Type และ (2) กำหนด Missing Values



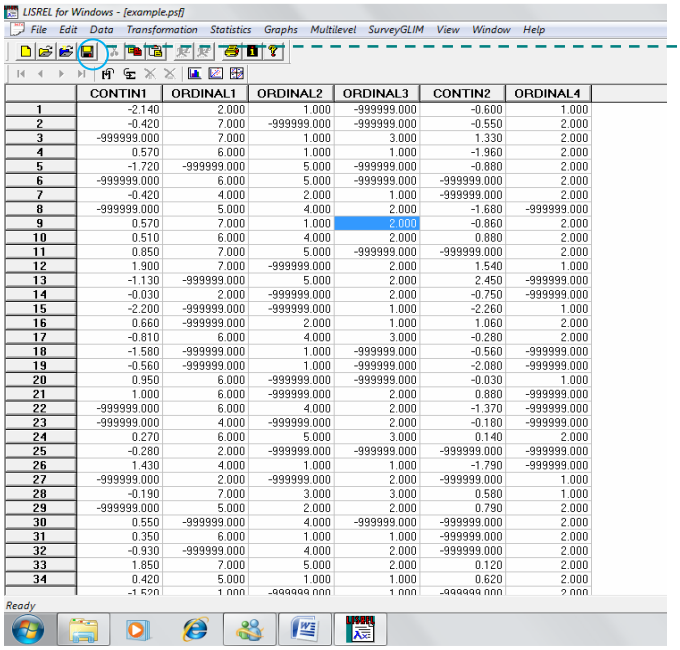
!!!ข้อควรระวังในการกำหนดตัวแปร เลือกมาตรวัดให้ตรงกับลักษณะของข้อมูลว่าเป็น Ordinal หรือ Continuous

!!!Click “Apply to all” กรณีต้องการให้มีผลต่อตัวแปรอื่นๆ ทั้งหมด



!!!การกำหนดค่า Missing สามารถระบุได้ทั้งหมด 3 ค่า หรือระบุเป็นช่วงก็ได้ Low - High

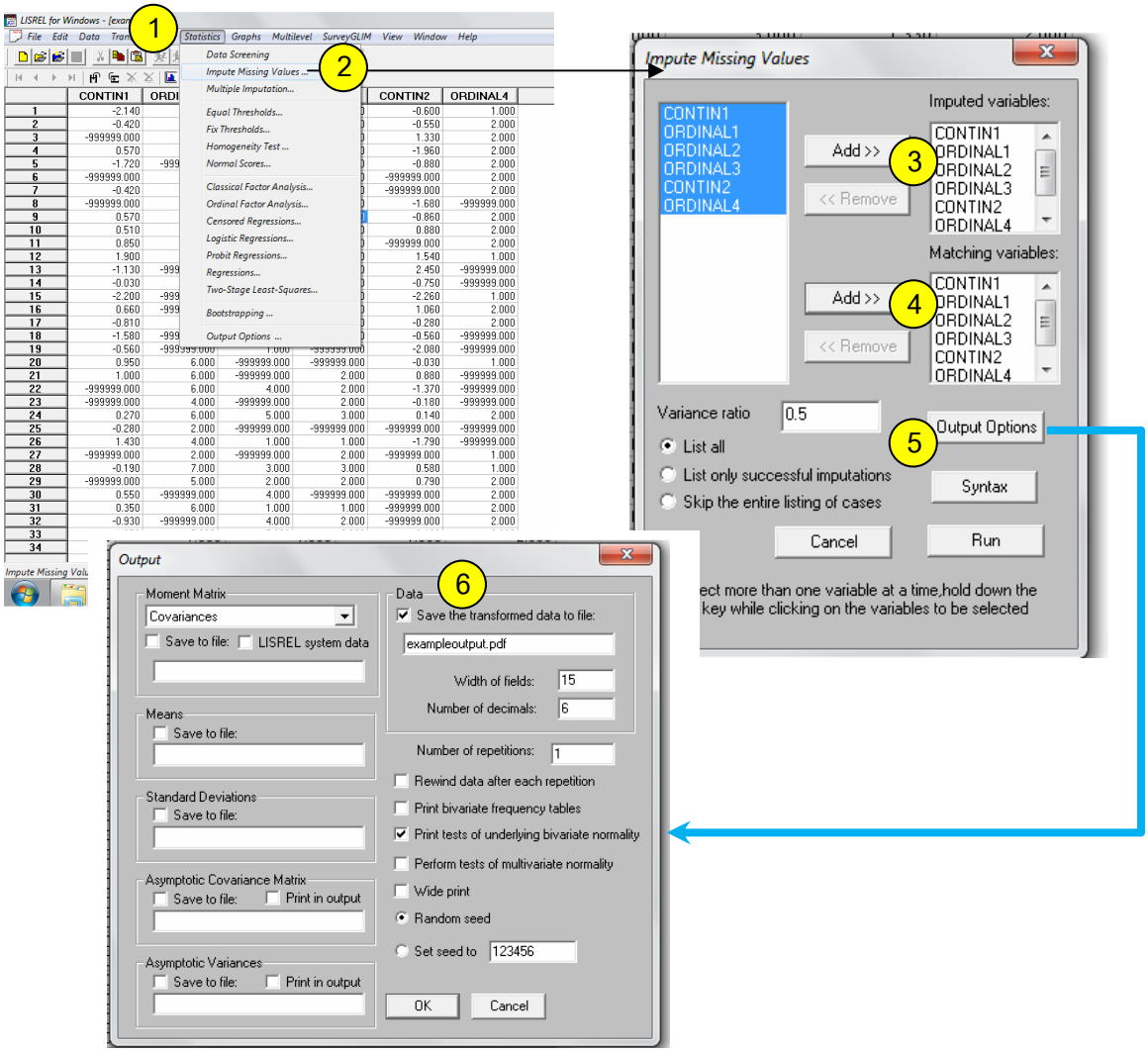
!!!ระวัง....หลัง Define missing values เรียบร้อยแล้ว ให้บันทึกข้อมูลก่อนทุกครั้ง โดยให้สังเกตที่ Icon บันทึกข้อมูลว่า Active หรือไม่ (เมื่อท่านบันทึกแล้ว Icon Save จะไม่ Active



บันทึกแล้ว Icon Save จะไม่ Active

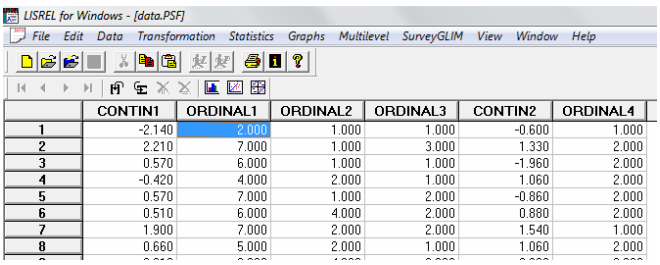
5. การ Impute ข้อมูลที่สูญหาย

เลือกเมนู (1) Statistics ----> เลือก (2) Impute Missing Values----->เลือก (3) ตัวแปร ------> เลือก (4) ตัวแปรที่จะจับคู่ (เลือกตัวแปรใดตัวแปรหนึ่งหรือหลายๆตัวแปรก็ได้) ----->เลือก (5) Output Option-----> Click (6) "Save" และตั้งชื่อไฟล์นามสกุล ".psf"---->กด OK -----> Click "Run"



6. **ผลลัพธ์** (Output)

ไฟล์ใหม่ที่เราบันทึกเมื่อเปิดดูจะพบว่าข้อมูลที่สูญหาย (Missing) นั้นได้ถูกแทนที่ด้วยค่าที่ผ่านกระบวนการ Impute แบบ Matching เรียบร้อยแล้ว



	CONTIN1	ORDINAL1	ORDINAL2	ORDINAL3	CONTIN2	ORDINAL4
1	-2.140	2.000	1.000	1.000	-0.600	1.000
2	2.210	7.000	1.000	3.000	1.330	2.000
3	0.570	6.000	1.000	1.000	-1.960	2.000
4	-0.420	4.000	2.000	1.000	1.060	2.000
5	0.570	7.000	1.000	2.000	-0.860	2.000
6	0.510	6.000	4.000	2.000	0.880	2.000
7	1.900	7.000	2.000	2.000	1.540	1.000
8	0.660	5.000	2.000	1.000	1.060	2.000

หน้าต่างไฟล์ใหม่ที่ค่าสูญหาย (Missing) ถูกแทนด้วยค่าใหม่เรียบร้อยแล้ว

- **!!!!เพิ่มเติม**

<u>วิธีการจัดการกับข้อมูลสูญหาย</u> (Methods of handling missing data)

การจัดการกับข้อมูลสูญหายมีหลายวิธีการให้เลือกใช้ การพิจารณาเลือกใช้วิธีการใดขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้น หากเลือกวิธีการที่ไม่เหมาะสมมาใช้อาจเป็นการเพิ่มค่าความคลาดเคลื่อนและทำลายผลลัพธ์ที่ควรจะได้ สำหรับวิธีการจัดการกับข้อมูลสูญหายที่มักถูกเลือกนำมาใช้ดังนี้

1. Listwise data deletion เป็นวิธีการจัดการกับข้อมูลสูญหายที่ง่ายมาก นั่นคือไม่สนใจข้อมูลสูญหายที่เกิดขึ้น โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลเฉพาะส่วนที่สมบูรณ์ แนวทางนี้จะมีความเหมาะสมในกรณีที่ข้อมูลสูญหายมีจำนวนน้อยมาก และ/หรือผลจากการวิเคราะห์มีความชัดเจนมาก วิธีการนี้มักถูกกำหนดให้ใช้เป็นหลัก (by default) สำหรับจัดการกับข้อมูลสูญหายในโปรแกรมคอมพิวเตอร์ทางสถิติทั่วๆไป หากไม่เจาะจงเลือกใช้วิธีการอื่นในการจัดการกับข้อมูลสูญหาย
2. Pairwise data deletion สำหรับกรณีที่ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ โดยจะทำการวิเคราะห์ข้อจากข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปร
3. Mean substitution เป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าในแต่ละกลุ่มย่อยของตัวแปรอื่น ซึ่งเป็นวิธีที่พัฒนามาจากการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่า ทั้งนี้เนื่องจากข้อสมมติที่ว่าค่าของข้อมูลสูญหายควรรจะต้องขึ้นอยู่กับลักษณะของหน่วยตัวอย่าง โดยลักษณะของหน่วยตัวอย่างที่ใกล้เคียงกันควรจะมีค่าข้อมูลที่สนใจคล้ายคลึงกัน
4. Regression Method ทำการสร้างสมการถดถอยระหว่างตัวแปรใดๆ ที่ต้องการจากข้อมูลที่สมบูรณ์ โดยกำหนดให้ตัวแปรตามเป็นตัวแปรที่มีข้อมูลไม่สมบูรณ์ จากนั้นใช้สมการถดถอยที่ได้ทำการประมาณค่าของข้อมูลที่ไม่สมบูรณ์
5. Hot deck imputationเป็นวิธีการพิจารณาเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกันมากที่สุดกับหน่วยตัวอย่างที่เกิดค่าสูญหาย จากนั้นแทนค่าที่สูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงนั้น
6. Expectation Maximization (EM) approachวิธีการนี้เป็นการอาศัยหลักของกระบวนการวนซ้ำ (iterative procedure) ระหว่าง 2 ขั้นตอน โดยขั้นตอนแรก เป็นขั้นตอนที่เรียกว่า Expectation (E) step ซึ่งจะทำการประมาณค่าคาดหวังจากฟังก์ชัน likelihood ภายใต้ข้อมูลที่สมบูรณ์ สำหรับขั้นตอนที่สอง เป็นขั้นตอนที่เรียกว่า Maximization (M) step เพื่อทำการแทนค่าคาดหวังของข้อมูลสูญหายด้วยค่าที่ได้จาก E step และทำการประมาณค่าคาดหวังจากฟังก์ชัน likelihood ในกรณีถ้าไม่เกิดข้อมูลสูญหาย โดยจะทำการวนซ้ำระหว่าง 2 ขั้นตอนจนกว่าจะเกิดค่าที่ลู่เข้า (convergence) หรือค่าที่มีการเปลี่ยนแปลงน้อยมาก ใช้ค่านั้นแทนค่าข้อมูลสูญหายที่เกิดขึ้น
7. Raw maximum likelihood methodเป็นวิธีการที่อาศัยข้อมูลสมบูรณ์ในการสร้างค่า maximum likelihood ภายใต้ตัวแบบทางสถิติที่เหมาะสม ไม่ว่าจะเป็น structural equation model, regression model, ANOVA และ ANCOVA models
8. Multiple imputation (MI)เป็นวิธีการที่ผสมผสานระหว่างวิธีการ EM และ Raw maximum likelihood methods ร่วมกับความสามารถของคุณสมบัติ hot deck เพื่อทำการสร้างชุดจำลองของข้อมูลที่ได้ทำ

การแทนค่าข้อมูลสูญหายด้วย imputed value แล้วขึ้นมาหลาย ๆ ชุด (ประมาณ 5 ถึง 10 ชุด) จากนั้นทำการวิเคราะห์ข้อมูลจากชุดต่าง ๆ บันทึกผลการวิเคราะห์ที่ได้ โดยผลการวิเคราะห์ที่ได้เหล่านี้จะถูกรวมเข้าด้วยกันเพื่อทำการ

บรรณานุกรม

.....หนังสือประกอบการเรียน

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี. (2006) **ข้อมูลสูญหายและแนวทางการจัดการ**. Data Management & Biostatistics Journal. Volume 4.