การจัดการข้อมูลสูญหาย

สันทัด พรประเสริฐมานิต

โครงร่างในการนำเสนอ

- กระบวนการเกิดข้อมูลสูญหาย
- วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย
- วิธีการใหม่ในการจัดการข้อมูลสูญหาย
- ตัวแปรช่วยทำนายข้อมูลสูญหาย
- การออกแบบจงใจให้มีข้อมูลสูญหาย

การจัดการข้อมูลสูญหาย

- ในการเก็บข้อมูล หากผู้วิจัยไม่สามารถเก็บข้อมูลทั้งหมดได้ จะทำให้เกิดข้อมูลสูญหาย (Missing data)
- ข้อมูลสูญหาย สามารถเกิดได้หลายสาเหตุ
  - กลุ่มตัวอย่างไม่อยากตอบคำถามบางข้อ เช่น รายได้
  - กลุ่มตัวอย่างข้ามคำถามบางข้อโดยบังเอิญ
  - กลุ่มตัวอย่างไม่สามารถตอบคำถามบางข้อได้ เช่น เพศชายไม่สามารถตอบคำถามเกี่ยวกับการตั้งครรภ์ได้

รูปแบบข้อมูลสูญหาย

- รูปแบบข้อมูลสูญหาย (Missing value patterns)

X	Y_1	Y_2	Y_3	Y_4	Z
5	7	3	1	3	5
4	8	?	5	4	4
3	?	2	7	5	1
4	7	2	7	3	2
6	?	?	?	2	2
3	5	7	3	6	3

รูปแบบข้อมูลสูญหาย 4 กลุ่ม

1. ไม่มีข้อมูลสูญหาย (4 คน)
2. มีข้อมูลสูญหายที่ Y_1 (1 คน)
3. มีข้อมูลสูญหายที่ Y_2 (1 คน)
4. มีข้อมูลสูญหายที่ Y_1, Y_2, Y_3 (1 คน)

รูปแบบข้อมูลสูญหาย

- แม้ว่ารูปแบบข้อมูลสูญหายเป็นอย่างไร การจัดการข้อมูลสูญหายก็ใกล้เคียงกัน แต่คุณภาพของผลที่ได้อาจต่างกัน ขึ้นอยู่กับความสมบูรณ์ของข้อมูล

Y_1	Y_2	Y_3
7	?	1
8	?	5
5	?	7
?	2	7
?	5	5
?	7	3

ข้อมูลในด้านใด ทำนายความสัมพันธ์ระหว่าง Y_1 และ Y_2 ได้ดีกว่ากัน

Y_1	Y_2	Y_3
?	?	1
?	?	5
?	?	7
7	2	7
4	5	5
5	7	3

รูปแบบข้อมูลสูญหาย

- แม้ว่ารูปแบบข้อมูลสูญหายเป็นอย่างไร การจัดการข้อมูลสูญหายก็ใกล้เคียงกัน แต่คุณภาพของผลที่ได้อาจต่างกัน ขึ้นอยู่กับความสมบูรณ์ของข้อมูล

Y_1	Y_2	Y_3
7	?	1
8	?	5
5	?	7
?	2	7
?	5	5
?	7	3

เราต้องการข้อมูล ที่ทำให้สามารถประมาณค่าความแปรปรวนร่วมระหว่างข้อมูลให้มากที่สุด หรือที่เรียกว่า ความครอบคลุมความแปรปรวนร่วม (Covariance Coverage)

Y_1	Y_2	Y_3
?	?	1
?	?	5
?	?	7
7	2	7
4	5	5
5	7	3

รูปแบบข้อมูลสูญหาย

- แม้ว่ารูปแบบข้อมูลสูญหายเป็นอย่างไร การจัดการข้อมูลสูญหายก็ใกล้เคียงกัน แต่คุณภาพของผลที่ได้อาจต่างกัน ขึ้นอยู่กับความสมบูรณ์ของข้อมูล

Y_1	Y_2	Y_3
7	?	1
8	?	5
5	?	7
?	2	7
?	5	5
?	7	3

ข้อมูลทางซ้าย เสียคุณภาพในการประมาณค่าความแปรปรวนร่วมระหว่าง Y_1 และ Y_2 มากกว่า

Y_1	Y_2	Y_3
?	?	1
?	?	5
?	?	7
7	2	7
4	5	5
5	7	3

รูปแบบข้อมูลสูญหาย

- แม้ว่ารูปแบบข้อมูลสูญหายเป็นอย่างไร การจัดการข้อมูลสูญหายก็ใกล้เคียงกัน แต่คุณภาพของผลที่ได้อาจต่างกัน ขึ้นอยู่กับความสมบูรณ์ของข้อมูล

Y_1	Y_2	Y_3
7	?	1
8	?	5
5	?	7
?	2	7
?	5	5
?	7	3

สัดส่วนข้อมูลที่เสียไปจากค่าสูญหาย (Fraction Missing Information; FMI)

ยิ่งมาก ยิ่งแสดงว่าการฟื้นคืนข้อมูลที่เสียไปจากข้อมูลสูญหาย ยิ่งไม่ดี

FMI ของ $r(Y_1, Y_2) = 1.0$

FMI ของ $r(Y_1, Y_2) = 0.5$

Y_1	Y_2	Y_3
?	?	1
?	?	5
?	?	7
7	2	7
4	5	5
5	7	3

กระบวนการเกิดข้อมูลสูญหาย

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

กระบวนการเกิดข้อมูลสูญหาย

- Missing Completely at Random (MCAR) การเกิดข้อมูลสูญหาย ไม่ได้เกิดอย่างเป็นระบบ เช่น กลุ่มตัวอย่างลืมตอบคำถาม

เพศ	น้ำหนัก
ชาย	75
หญิง	45
ชาย	?
ชาย	65
หญิง	55
หญิง	50

ผู้ชายคนนี้ลืมตอบคำถามพอดี

กระบวนการเกิดข้อมูลสูญหาย

- Missing at Random (MAR) การเกิดข้อมูลสูญหาย ขึ้นอยู่กับตัวแปรหนึ่งที่ได้เก็บข้อมูลเอาไว้ หากใช้วิธีการจัดการที่ถูกต้อง จะทำให้ผลการวิเคราะห์ไม่เพี้ยนไป เช่น โอกาสในการตอบอายุ ขึ้นอยู่กับเพศ (ซึ่งได้เก็บข้อมูลเอาไว้)

เพศ	น้ำหนัก
ชาย	75
หญิง	45
ชาย	70
ชาย	65
หญิง	?
หญิง	?

ผู้หญิงมีแนวโน้มไม่ตอบน้ำหนักมากกว่า

กระบวนการเกิดข้อมูลสูญหาย

- Missing Not at Random (MNAR) การเกิดข้อมูลสูญหาย ขึ้นอยู่กับตัวแปรหนึ่งที่ไม่ได้เก็บข้อมูลเอาไว้ ทำให้ไม่สามารถจัดการได้ง่าย (ยกเว้นแต่ใช้เทคนิคบางอย่าง ซึ่งมีข้อตกลงเบื้องต้นมากมาย) เช่น โอกาสในการตอบรายได้ ขึ้นอยู่กับรายได้เอง

เพศ	น้ำหนัก
ชาย	?
หญิง	45
ชาย	70
ชาย	65
หญิง	50
หญิง	?

คนน้ำหนักเยอะ มีแนวโน้มไม่ตอบคำถามนี้

กระบวนการเกิดข้อมูลสูญหาย

- ในข้อมูลหนึ่ง ตัวแปรแต่ละตัว อาจมีกระบวนการเกิดข้อมูลสูญหายแตกต่างกัน หรือตัวแปรเดียวกัน อาจมีกระบวนการเกิดข้อมูลสูญหายหลายอย่างพร้อมกัน
- ไม่มีตัวแปรใดที่สามารถตัดสินได้ว่าเป็น MAR และ MNAR อย่างชัดเจน อาจมองได้ว่าข้อมูลสูญหายนั้น ถูกอธิบายด้วยตัวแปรอื่นที่เก็บข้อมูลมา มากน้อยเพียงใด ถ้ามากก็จะใกล้เคียง MAR ถ้าน้อยก็จะใกล้เคียง MNAR
  - เช่น ผู้หญิงที่น้ำหนักมาก มีแนวโน้มไม่ตอบน้ำหนักของตนเอง

กระบวนการเกิดข้อมูลสูญหาย

- MCAR และ MAR สามารถจัดการได้ด้วยวิธีที่จะอธิบายในที่นี้
- MNAR แก้ไขได้ยากมาก วิธีการที่ดีที่สุด คือ พยายามเปลี่ยนแปลงให้ใกล้เคียง MAR มากที่สุด โดยเก็บข้อมูลทำนายค่าสูญหายให้มากที่สุด
  - เช่น ไม่ทราบน้ำหนักของบางคน อาจเก็บข้อมูลความสูง นิสัยการกิน เพื่อทำนายน้ำหนักที่สูญหายได้ดีที่สุด
- ตัวแปรที่สามารถทำนายค่าสูญหายได้ แต่ไม่เกี่ยวข้องกับสิ่งที่ผู้วิจัยสนใจ จะเรียกว่า ตัวแปรช่วยทำนายข้อมูลสูญหาย (Auxiliary variables)

กระบวนการเกิดข้อมูลสูญหาย

- บางโปรแกรมมีการทดสอบ MCAR ขึ้น เช่น วิธีการทำ *t*-test หรือ Little's MCAR test
- แท้จริงแล้วก็คือ การทดสอบว่าค่าสูญหายไม่สัมพันธ์กับข้อมูลที่เก็บได้ (โดยไม่สนใจข้อมูลที่ไม่ได้เก็บ)
- ดังนั้น จึงบอกได้เพียงว่า ข้อมูลใกล้เคียงกับ MCAR หรือ MAR มากกว่า โดยที่ไม่ได้ตัดความเป็นไปได้ที่จะเป็น MNAR
- นอกจากนี้ วิธีการจัดการข้อมูลสูญหายที่ควรใช้กับ MCAR และ MAR ก็เหมือนกัน ดังนั้นไม่จำเป็นต้องทดสอบ MCAR

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- การกำจัดข้อมูล (Deletion methods) เช่น Listwise deletion; Pairwise deletion
- การแทนค่าข้อมูล 1 ครั้ง (Single imputation techniques) เช่น Mean imputation, Regression imputation, Average available items

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- Listwise deletion คือ หากข้อมูลของคนใด มีข้อมูลสูญหาย จะลบข้อมูลของคนนั้นทิ้งออกไปเลย

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

เช่น วิเคราะห์ถดถอย โดยใช้ข้อมูลแค่ 4 คน

ตัดข้อมูลบางส่วนทิ้ง ทั้งที่ควรจะใช้

Standard error น้อยลงกว่าความเป็นจริง

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- Pairwise deletion คือ การวิเคราะห์ข้อมูลที่ใช้ข้อมูล 2 ตัวแปร จะใช้ข้อมูลคู่ที่มีทั้งหมด

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

หาความสัมพันธ์ระหว่าง

- *X* กับ *Y* ใช้ข้อมูล 4 คน
- *X* กับ *Z* ใช้ข้อมูล 5 คน
- *Y* กับ *Z* ใช้ข้อมูล 4 คน

ข้อมูลทุกคู่ไม่เท่าเทียมกัน อาจทำให้เกิดเมทริกซ์สหสัมพันธ์ที่เป็นไปไม่ได้ในทางสถิติ

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- การกำจัดข้อมูล เป็นการคัดข้อมูลทิ้งทั้งหมด หากในแถวนั้นมีข้อมูลสูญหายในตัวแปรอื่น
- การกำจัดข้อมูลจึงทำให้กำลังในการทดสอบทางสถิติน้อยลง (และ SE สูงขึ้น) กว่าที่ควรจะเป็น เนื่องจากไม่ได้ใช้ข้อมูลที่มีทั้งหมด
- อาจทำให้การประมาณค่าพารามิเตอร์ผิดพลาดมาก โดยเฉพาะอย่างยิ่งในข้อมูลแบบ MAR หรือ MNAR
- หมายเหตุ Listwise deletion เป็นวิธีการที่โปรแกรมทางสถิติเกือบทุกโปรแกรม ใช้เป็นค่าเริ่มต้นในการจัดการข้อมูลสูญหาย

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- Mean imputation คือ การแทนข้อมูลสูญหายของตัวแปรนั้น ด้วยค่าเฉลี่ยของตัวแปร

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

แทนค่าด้วย (7 + 8 + 7 + 9) / 4 = 7.75

ทำให้ความแปรปรวนต่ำกว่าความเป็นจริง และความสัมพันธ์ระหว่างตัวแปรน้อยกว่าความเป็นจริง

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- Regression method คือ การแทนค่าข้อมูลสูญหาย จากการทำนายผ่าน Regression ด้วยตัวแปรอื่นในข้อมูล

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

$$\hat{Y} = 5.611 + 0.583X - 0.194Z$$

$$\hat{Y} = 5.611 + 0.583(3) - 0.194(1)$$
$$= 7.17$$

ทำให้ความสัมพันธ์ระหว่างตัวแปรสูงกว่าความเป็นจริง

วิธีการดั้งเดิมในการจัดการข้อมูลสูญหาย

- Averaging Available Items คือ การแทนค่าข้อคำถามที่สูญหาย จากคะแนนเฉลี่ยของข้อคำถามอื่นจากกลุ่มตัวอย่างเดียวกัน

X	Y_1	Y_2	Y_3	Y_4	Z
5	7	3	1	3	5
4	8	9	5	4	4
3	?	2	7	5	1
4	7	2	7	3	2
6	9	8	5	2	2

แทนค่าด้วย (2 + 7 + 5) / 3 = 4.67

ข้อตกลงเบื้องต้น คือ ข้อคำถามต้องเป็นแบบคู่ขนาน (Parallel) ซึ่งมักไม่เป็นจริง

วิธีการดั้งเดิมการจัดการข้อมูลสูญหาย

- การแทนค่าข้อมูล 1 ครั้ง เป็นการเติมข้อมูลที่น่าจะใกล้เคียงกับความเป็นจริงมากที่สุดลงในข้อมูลสูญหาย
- บางวิธีอาจทำให้ความแปรปรวนของข้อมูลอาจลดลงกว่าที่ควรจะเป็น
- บางวิธีอาจทำให้ความสัมพันธ์เพิ่มขึ้นกว่าที่ควรจะเป็น
- วิธีเหล่านี้ ไม่ได้คำนึงว่า ข้อมูลที่แทนค่านั้น ไม่ได้เป็นข้อมูลที่สังเกตได้ แต่เป็นข้อมูลที่ประมาณค่าขึ้นมา ที่ต้องคำนึงถึงความไม่แน่นอนของข้อมูลด้วย

วิธีการใหม่การจัดการข้อมูลสูญหาย

- การแทนค่าหลายครั้ง (Multiple imputation)
- การวิเคราะห์ความเป็นไปได้สูงสุดโดยตรง (Direct maximum likelihood)

วิธีการใหม่การจัดการข้อมูลสูญหาย

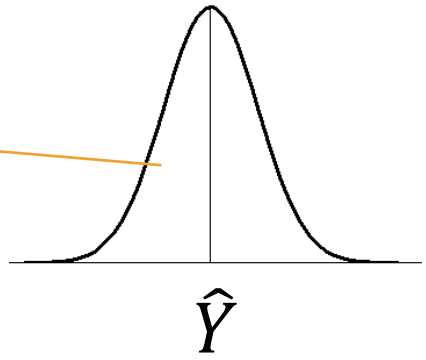
- Multiple imputation คือ การแทนค่าข้อมูลสูญหายรูปแบบหนึ่ง ซึ่งจะแทนค่าข้อมูลสูญหาย ด้วยค่าที่ทำนายได้บวกกับความผิดพลาดในการทำนาย
- เพื่อไม่ให้ผลการวิเคราะห์ได้รับอิทธิพลจากความผิดพลาดในการทำนายที่สุ่มออกมาเพียงแค่ตัวเดียว นักวิจัยจึงแทนค่าออกมาหลายครั้ง ได้ข้อมูลหลายชุด
- นำข้อมูลแต่ละชุดมาวิเคราะห์สถิติที่ต้องการ ได้ค่าสถิติที่ต้องการหลายตัว
- นำค่าสถิติที่ได้ทั้งหมด มารวมกัน และหา Standard error ที่ถูกต้อง ด้วยวิธีการของ Rubin

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

$$\hat{Y} = 5.611 + 0.583(3) - 0.194(1)$$
$$= 7.17$$

สมมติว่า $SE(\hat{Y}) = 2$

สุ่มค่าจากโค้งปกติ



X	Y	Z
5	7	5
4	8	4
3	10.02	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	3.48	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	6.15	1
4	7	2
6	9	2

X	Y	Z
5	7	5
4	8	4
3	10.02	1
4	7	2
6	9	2

$\hat{Y} = 9.8 - 0.03X - 0.53Z$

X	Y	Z
5	7	5
4	8	4
3	3.48	1
4	7	2
6	9	2

$\hat{Y} = 0.2 + 1.37X + 0.24Z$

X	Y	Z
5	7	5
4	8	4
3	6.15	1
4	7	2
6	9	2

$\hat{Y} = 4.1 + 0.8X - 0.1Z$

ข้อมูลแทนค่า	b_1	$SE(b_1)$
1	-0.028	0.631
2	1.374	0.752
3	0.801	0.425

ใช้วิธีการรวมข้อมูลของ Rubin

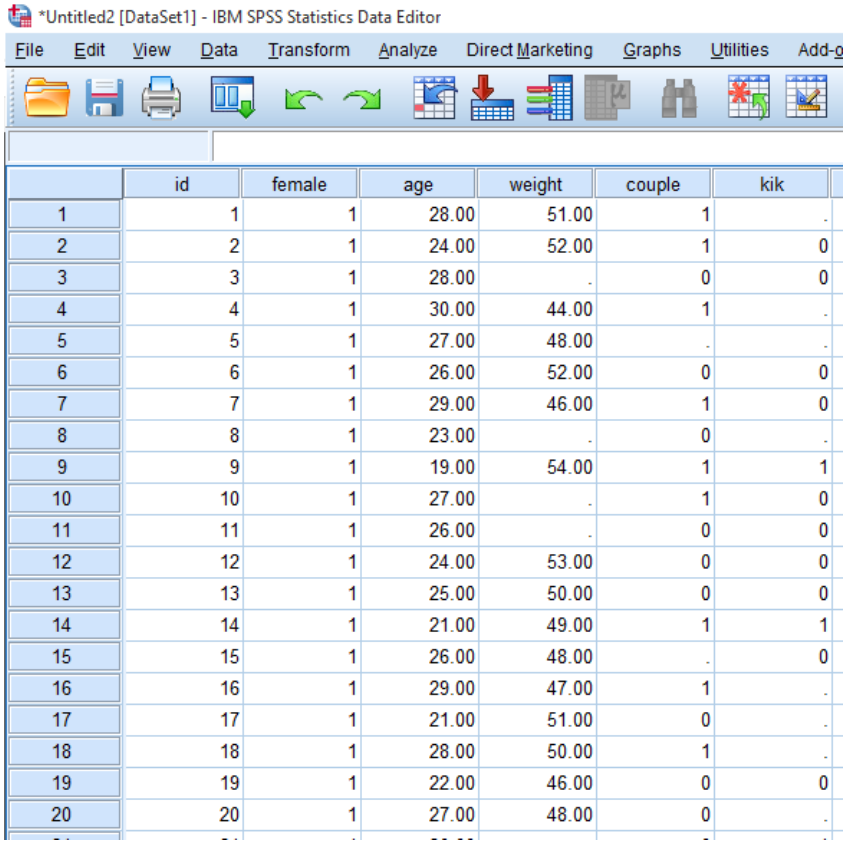
$b_1 = 0.715$

$SE(b_1) = 1.022$

วิธีการใหม่การจัดการข้อมูลสูญหาย

- ในกรณีที่ตัวแปรหนึ่ง มีข้อมูลสูญหายไม่ถึง 5% การเลือกวิธีการดั้งเดิมไม่ได้มีผลกระทบมาก แต่ถ้าใช้วิธีการใหม่ได้ ให้ใช้ดีกว่า
- แต่หากตัวแปรใด มีข้อมูลสูญหายเกิน 5% ควรใช้วิธีการใหม่
- SPSS สามารถจัดการข้อมูลสูญหายได้ด้วยวิธี Multiple imputation (MI)
- วิธี Multiple imputation ถือเป็นวิธีการเตรียมข้อมูล การนำไปวิเคราะห์จริง การแทนค่าเพียงครั้งเดียว สามารถนำไปใช้ในการวิเคราะห์หลายครั้งได้

การแทนค่าแบบพหุ



	id	female	age	weight	couple	kik
1	1	1	28.00	51.00	1	.
2	2	1	24.00	52.00	1	0
3	3	1	28.00	.	0	0
4	4	1	30.00	44.00	1	.
5	5	1	27.00	48.00	.	.
6	6	1	26.00	52.00	0	0
7	7	1	29.00	46.00	1	0
8	8	1	23.00	.	0	.
9	9	1	19.00	54.00	1	1
10	10	1	27.00	.	1	0
11	11	1	26.00	.	0	0
12	12	1	24.00	53.00	0	0
13	13	1	25.00	50.00	0	0
14	14	1	21.00	49.00	1	1
15	15	1	26.00	48.00	.	0
16	16	1	29.00	47.00	1	.
17	17	1	21.00	51.00	0	.
18	18	1	28.00	50.00	1	.
19	19	1	22.00	46.00	0	0
20	20	1	27.00	48.00	0	.

ตัวอย่างข้อมูล

Female: 1 = หญิง, 0 = ชาย

Age: อายุ

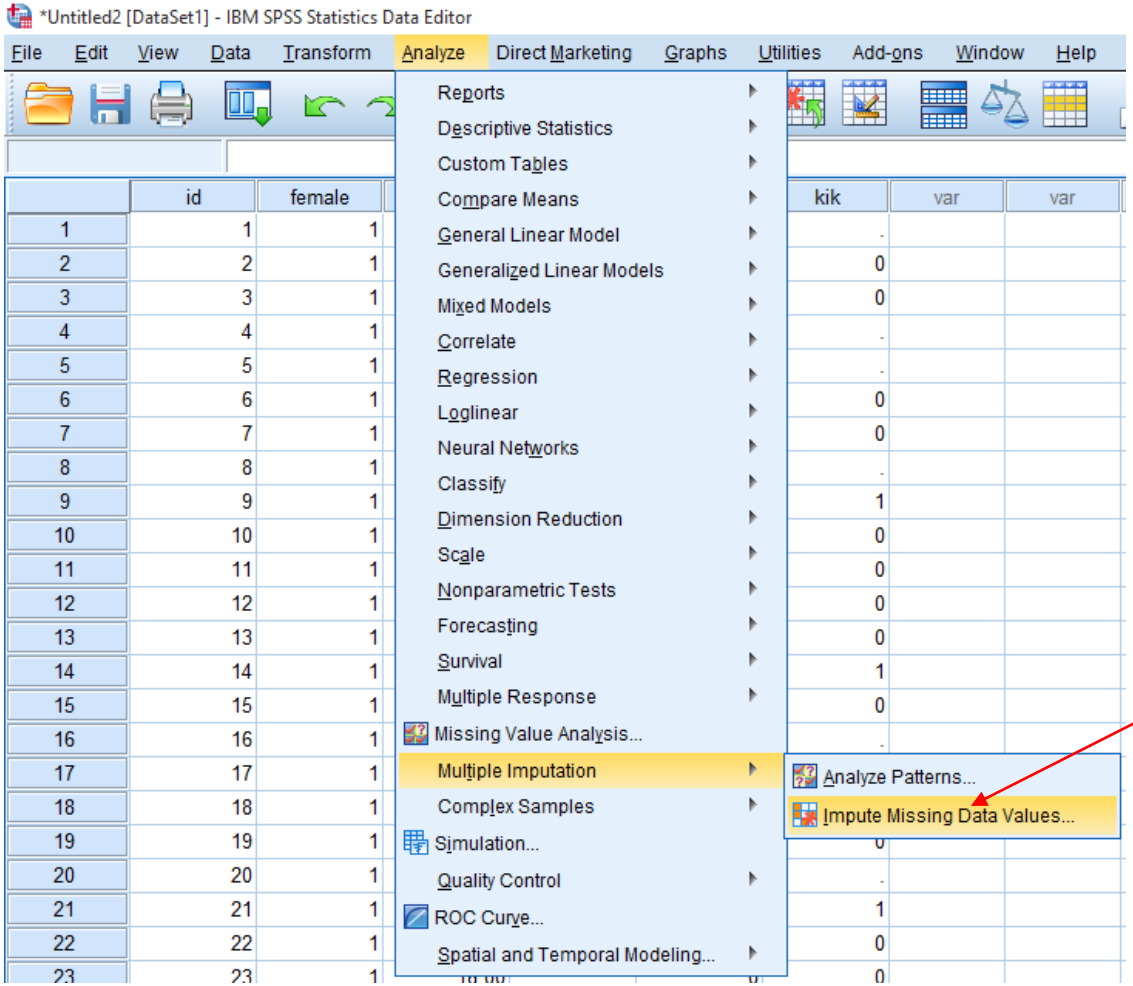
Weight: น้ำหนัก

Couple: มีแฟนอย่างเป็นทางการ

Kik: มีกิ๊ก

หมายเหตุ ตัวแปรจะต้องกำหนดระดับการวัดของแต่ละตัวแปรใน Variable View… ให้ถูกต้อง (Nominal/Ordinal/Scale)

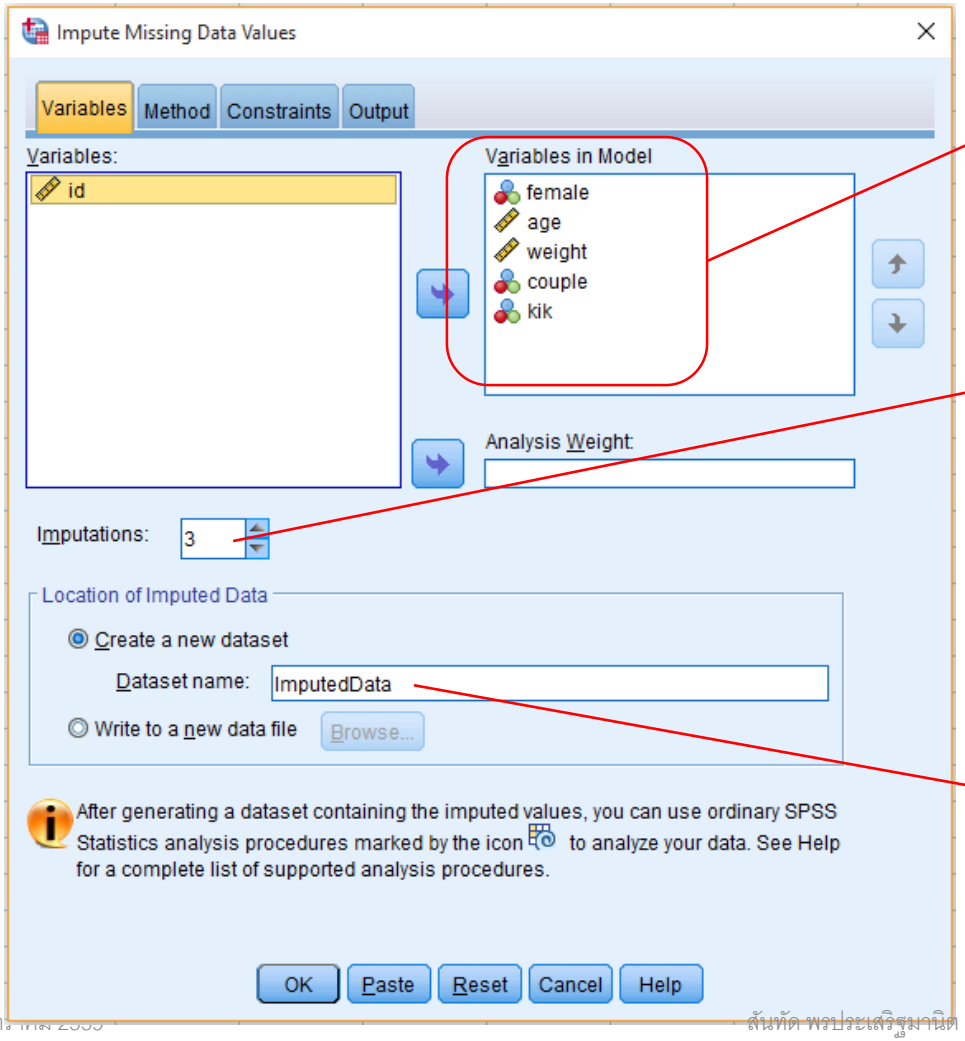
การแทนค่าแบบพหุ



เลือกเพื่อทำ

Multiple Imputation

การแทนค่าแบบพหุ

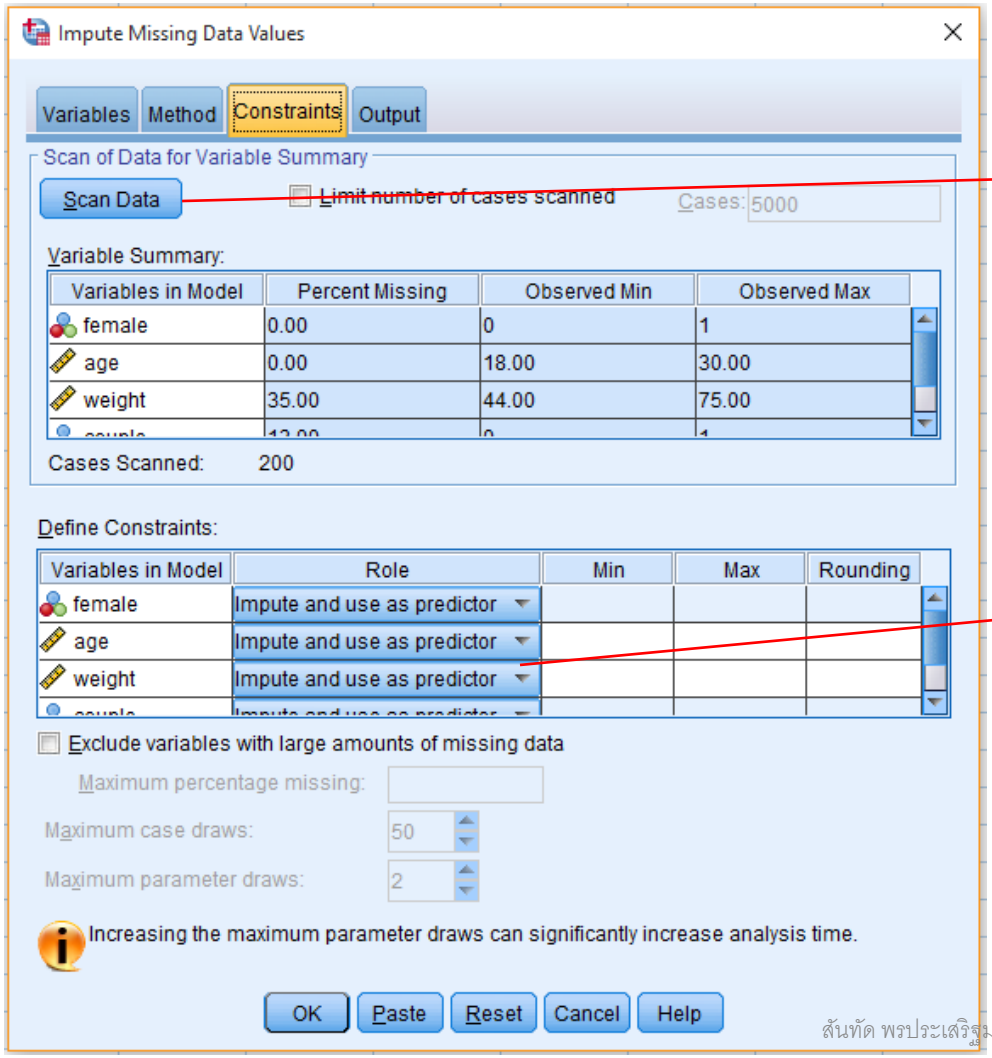


เลือกตัวแปรที่ต้องการแทนค่าสูญหายหรือตัวแปรที่ใช้ในการทำนายค่าสูญหายของตัวแปรอื่น

จำนวนครั้งในการแทนค่าสูญหาย ยิ่งเยอะยิ่งดี แต่จะเสียเวลาในการวิเคราะห์ ผมแนะนำให้ใช้ประมาณ 20 ครั้ง

ใส่ชื่อข้อมูลที่แทนค่าแล้ว

การแทนค่าแบบพหุ



สามารถกด Scan Data เพื่อตรวจสอบลักษณะของข้อมูลโดยสังเขป

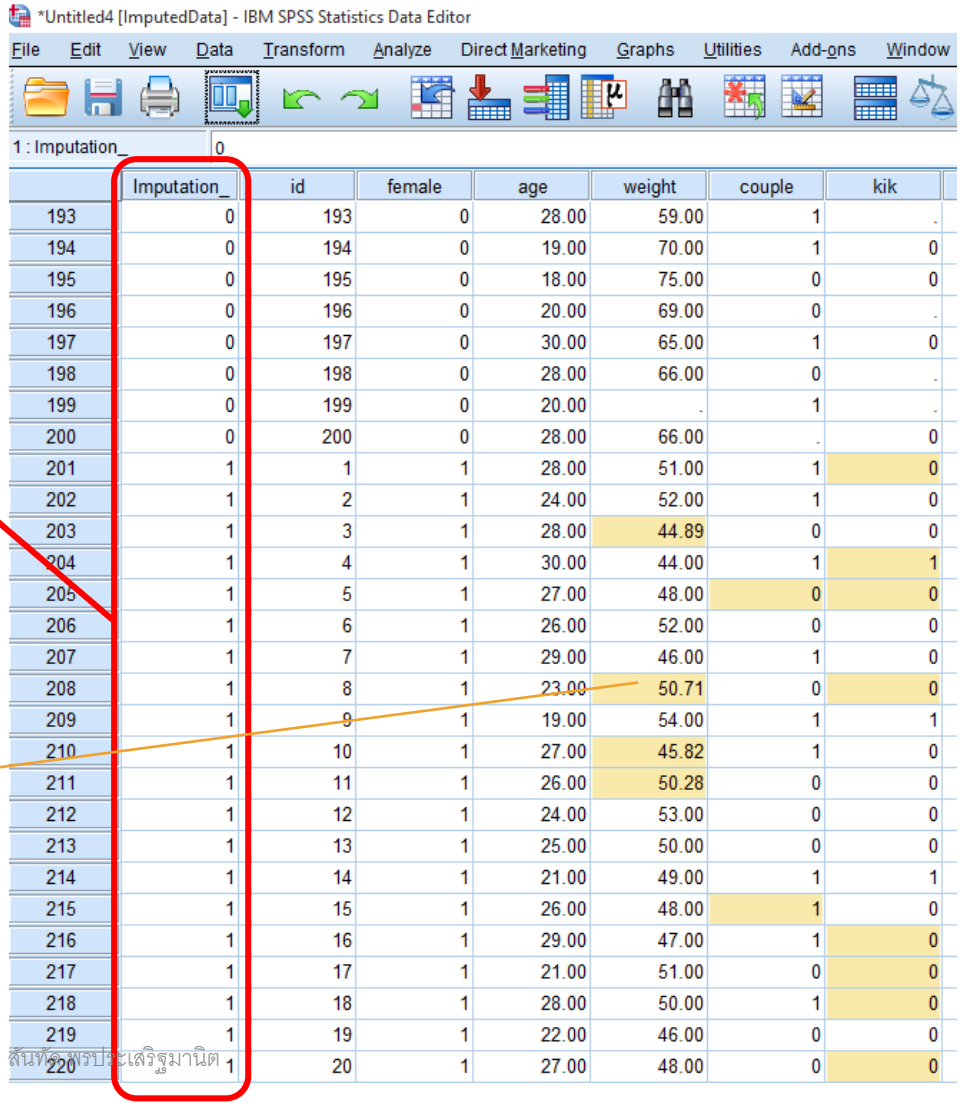
สามารถเลือกได้ว่าตัวแปรใดให้มีการแทนค่า ตัวแปรใดใช้ทำนายค่าสูญหายของตัวแปรอื่น

การแทนค่าแบบพหุ

ข้อมูลใหม่หลังจากแทนค่าแล้ว

ลำดับของครั้งที่แทนค่า
0 คือข้อมูลเดิม

สีเหลือง คือ ค่าที่โปรแกรมแทนค่าสูญหาย



	Imputation_	id	female	age	weight	couple	kik
193	0	193	0	28.00	59.00	1	.
194	0	194	0	19.00	70.00	1	0
195	0	195	0	18.00	75.00	0	0
196	0	196	0	20.00	69.00	0	.
197	0	197	0	30.00	65.00	1	0
198	0	198	0	28.00	66.00	0	.
199	0	199	0	20.00	.	1	.
200	0	200	0	28.00	66.00	.	0
201	1	1	1	28.00	51.00	1	0
202	1	2	1	24.00	52.00	1	0
203	1	3	1	28.00	44.89	0	0
204	1	4	1	30.00	44.00	1	1
205	1	5	1	27.00	48.00	0	0
206	1	6	1	26.00	52.00	0	0
207	1	7	1	29.00	46.00	1	0
208	1	8	1	23.00	50.71	0	0
209	1	9	1	19.00	54.00	1	1
210	1	10	1	27.00	45.82	1	0
211	1	11	1	26.00	50.28	0	0
212	1	12	1	24.00	53.00	0	0
213	1	13	1	25.00	50.00	0	0
214	1	14	1	21.00	49.00	1	1
215	1	15	1	26.00	48.00	1	0
216	1	16	1	29.00	47.00	1	0
217	1	17	1	21.00	51.00	0	0
218	1	18	1	28.00	50.00	1	0
219	1	19	1	22.00	46.00	0	0
220	1	20	1	27.00	48.00	0	0

การแทนค่าแบบพหุ

- จากข้อมูลนี้ ลองทำนายน้ำหนัก ด้วยเพศและอายุ โดยการทำ Multiple Regression แบบปกติ
  - Analyze → Regression → Linear
  - ใส่ Female และ Age เป็นตัวแปรอิสระ และ Weight เป็นตัวแปรตาม

การแทนค่าแบบพหุ

Model Summary

Imputation Number	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Original data	1	.913[a]	.834	.831	3.33576
1	1	.915[a]	.836	.835	3.32228
2	1	.911[a]	.829	.827	3.39921
3	1	.915[a]	.837	.835	3.35190

a. Predictors: (Constant), age, female

ค่า R^2 ของข้อมูลดั้งเดิม
และข้อมูลที่แทนค่า

ANOVA[a]

Imputation Number	Model		Sum of Squares	df	Mean Square	F	Sig.
Original data	1	Regression	7081.360	2	3540.680	318.197	.000[b]
		Residual	1413.171	127	11.127		
		Total	8494.531	129			
1	1	Regression	11119.062	2	5559.531	503.692	.000[b]
		Residual	2174.401	197	11.038		
		Total	13293.463	199			
2	1	Regression	11047.687	2	5523.843	478.064	.000[b]
		Residual	2276.257	197	11.555		
		Total	13323.944	199			
3	1	Regression	11375.417	2	5687.709	506.240	.000[b]
		Residual	2213.335	197	11.235		
		Total	13588.752	199			

a. Dependent Variable: weight

b. Predictors: (Constant), age, female

การทดสอบ
$H_0: \mathrm{P}^2 = 0$
ของข้อมูลดั้งเดิม
และข้อมูลที่แทนค่า

การแทนค่าแบบพหุ

ผลการวิเคราะห์สัมประสิทธิ์ถดถอย ของข้อมูลดั้งเดิม และข้อมูลที่แทนค่า

Coefficients[a]

Imputation Number	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
			B	Std. Error	Beta					
Original data	1	(Constant)	67.352	1.914		35.181	.000			
		female	-14.989	.596	-.916	-25.150	.000			
		age	-.066	.075	-.032	-.884	.378			
1	1	(Constant)	68.701	1.511		45.463	.000			
		female	-14.960	.471	-.918	-31.730	.000			
		age	-.112	.060	-.054	-1.880	.062			
2	1	(Constant)	64.505	1.546		41.720	.000			
		female	-14.838	.482	-.909	-30.759	.000			
		age	.036	.061	.017	.582	.561			
3	1	(Constant)	66.314	1.525		43.495	.000			
		female	-15.096	.476	-.916	-31.735	.000			
		age	-.021	.060	-.010	-.349	.727			
Pooled	1	(Constant)	66.507	2.870		23.169	.000	.799	2.532	.790
		female	-14.965	.499		-29.971	.000	.096	.098	.969
		age	-.033	.105		-.310	.771	.759	2.036	.798

a. Dependent Variable: weight

ผลของการรวมค่าสถิติจากข้อมูลที่แทนค่าทั้งหมด

การแทนค่าแบบพหุ

- จากผลการวิเคราะห์ข้อมูล เพศมีผลต่อน้ำหนักอย่างมีนัยสำคัญ แต่อายุไม่มีผลอย่างมีนัยสำคัญ
- จะเห็นว่า การรวมผลการวิเคราะห์จะมีเฉพาะบางสถิติ เช่น สัมประสิทธิ์ถดถอย แต่สัมประสิทธิ์การทำนาย ไม่ได้ถูกรวมในที่นี้
- โปรแกรมอื่น มีความสามารถมากกว่า SPSS ในการทำ Multiple imputation เช่น Mplus, R

วิธีการใหม่การจัดการข้อมูลสูญหาย

- Direct maximum likelihood หรือ Full information maximum likelihood คือ การประมาณค่าทางสถิติ ผ่านข้อมูลทั้งหมดที่มี และข้ามข้อมูลที่ไม่มี วิธีการนี้จะต้องใช้โมเดลสมการเชิงโครงสร้าง (Structural equation modeling)

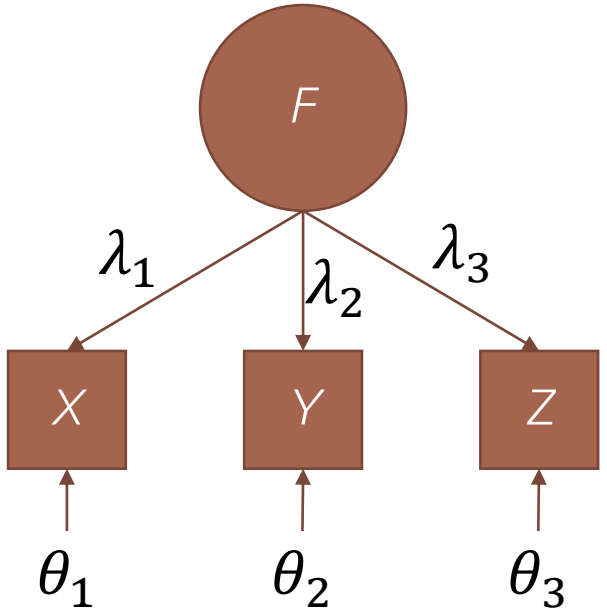
X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

ใช้ข้อมูลที่มีทั้งหมด ในการประมาณค่าทางสถิติ ซึ่ง 3 และ 1 ในข้อมูลของคนที่ *Y* สูญหายจะนำไปใช้คำนวณด้วย

วิธีการใหม่การจัดการข้อมูลสูญหาย

- เช่น จากข้อมูลนี้ ต้องการประมาณค่าผลการวิเคราะห์องค์ประกอบ

X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2



จากตรงนี้ Direct maximum likelihood จะต้องหาค่า $\lambda_1, \lambda_2, \lambda_3, \theta_1, \theta_2, \theta_3$ ที่มีแนวโน้มเป็นไปได้มากที่สุด ที่ทำให้เกิดข้อมูลชุดนี้ขึ้นมา (ข้ามค่าสูญหาย)

ตัวแปรช่วยทำนายข้อมูลสูญหาย

- เพื่อให้การวิเคราะห์ผลออกมาถูกต้อง จะต้องใช้วิธีการจัดการค่าสูญหายที่เหมาะสม โดยใช้วิธีการใหม่
- แต่วิธีการใหม่จะถูกต้องก็ต่อเมื่อ ค่าสูญหายเป็น MAR หรือ MCAR
- เพื่อลดโอกาสเกิดกระบวนการเกิดค่าสูญหายแบบ MNAR ผู้วิจัยจึงควรใส่ตัวแปรช่วยทำนายค่าสูญหาย (Auxiliary variable) แม้ผู้วิจัยจะไม่สนใจก็ตาม

ตัวแปรช่วยทำนายข้อมูลสูญหาย

- ใน Multiple imputation ผู้วิจัยเพียงแค่ใส่ตัวแปรช่วยเหล่านี้ ลงไปในโมเดลแทนค่า
- ใน Direct maximum likelihood ผู้วิจัยจะต้องสร้างโมเดลโดยเฉพาะ เพื่อให้ตัวแปรช่วยเหล่านี้ถูกประมาณค่าในโมเดล โดยไม่มีอิทธิพลต่อโมเดลหลัก เช่น Saturated Correlates
  - วิธีการนี้ ใน Mplus หรือ semTools package (ใน R) สามารถเพิ่มตัวแปรช่วย จากโมเดลที่สนใจโดยอัตโนมัติ

ตัวแปรช่วยทำนายข้อมูลสูญหาย

- หากมีตัวแปรช่วยมาก อาจทำให้การวิเคราะห์ข้อมูลนาน (อาจเป็นสัปดาห์)
- ดังนั้น จึงมีการเลือกตัวแปรช่วยทำนาย วิธีการหนึ่ง คือ หาความสัมพันธ์ระหว่างตัวแปรช่วยทำนาย กับการเกิดค่าสูญหายในตัวแปรที่สนใจ
  - หากมีค่าใดค่าหนึ่งสูงกว่า .4 ควรเก็บเอาไว้

เปรียบเทียบ

MULTIPLE IMPUTATION

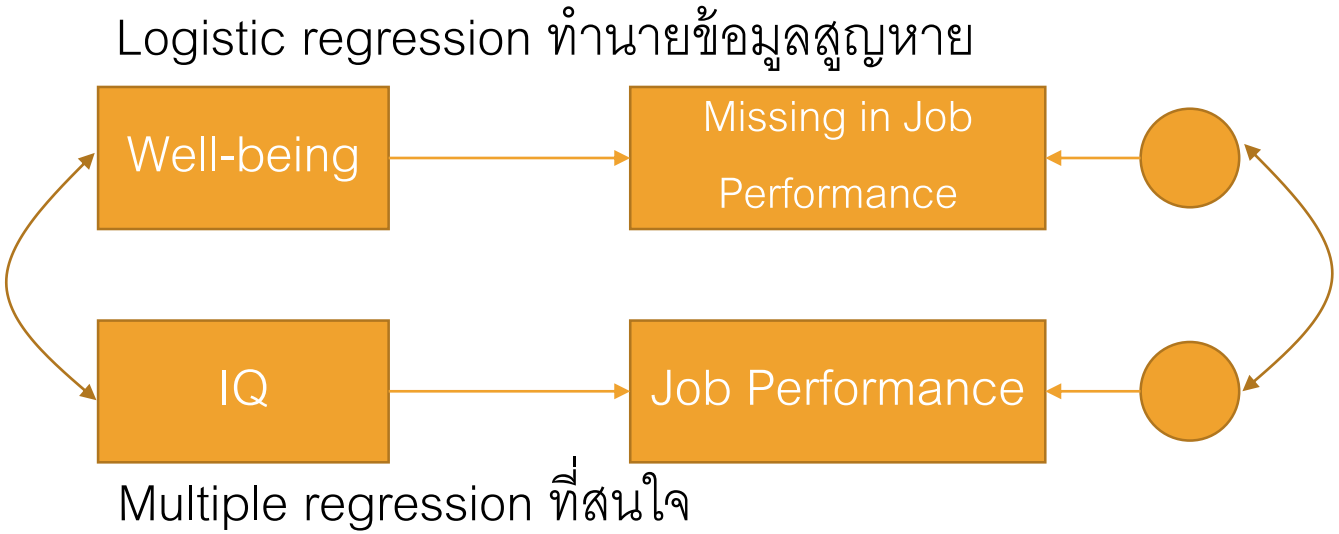
- การใส่ตัวแปรช่วยง่ายกว่ามาก
- สามารถจัดการข้อมูลสูญหายทีเดียว แล้ววิเคราะห์ข้อมูลหลายแบบได้
- สามารถจัดการข้อมูลแบบจัดกลุ่มได้
- สถิติบางชนิดไม่มีวิธีการรวม เช่น R^2

DIRECT MAXIMUM LIKELIHOOD

- สามารถจัดการข้อมูลที่มีปฏิสัมพันธ์ได้อย่างมีประสิทธิภาพ
- ใน SEM สามารถคำนวณ Fit indices ได้
- อาจมีปัญหาที่โมเดลไม่สามารถคำนวณได้ (Convergence problems)

วิธีการในการจัดการ MNAR

- Selection models เป็นการนำ Regression analysis มารวมกับโมเดลการทำนายค่าสูญหาย



วิธีการในการจัดการ MNAR

- Selection models จะสามารถประมาณค่าพารามิเตอร์ได้ถูกต้องเมื่อข้อตกลงเบื้องต้นถูกต้อง เช่น การกระจายของค่าคงเหลือ (Residuals) เป็น Bivariate normality ซึ่งไม่สามารถทดสอบได้
- หากข้อตกลงเบื้องต้นนี้ถูกละเมิด จะมีผลที่แย่กว่า Multiple imputation และ Direct maximum likelihood

วิธีการในการจัดการ MNAR

- Pattern mixture model เป็นการวิเคราะห์ข้อมูลแยกกัน ในแต่ละรูปแบบของค่าสูญหาย (Missing data patterns) แล้วผลการวิเคราะห์ข้อมูลจากแต่ละรูปแบบจะนำมารวมกัน

X	Y
5	78
8	87
9	65
7	70
6	?
5	?
6	?

เช่น ต้องการหาค่าเฉลี่ยของ Y

รูปแบบที่ 1: ข้อมูลสมบูรณ์ 4 คน

$$\bar{Y}_{(1)} = 86.6 - 1.6\bar{X}_{(1)}$$

รูปแบบที่ 2: Y สูญหาย 3 คน

$$\bar{Y}_{(2)} = \beta_{0(2)} + \beta_{1(2)}\bar{X}_{(2)}$$

ไม่สามารถประมาณค่าได้ เนื่องจากข้อมูลไม่เพียงพอ

วิธีการในการจัดการ MNAR

- แก้ไข โดยตั้งข้อตกลงเบื้องต้นว่าค่าสัมประสิทธิ์ถดถอยในรูปแบบที่ 2 เท่ากับรูปแบบที่ 1

X	Y
5	78
8	87
9	65
7	70
6	?
5	?
6	?

ใช้วิธีนี้วิเคราะห์ถดถอยอย่างง่าย

รูปแบบที่ 1: ข้อมูลสมบูรณ์ 4 คน

$$\bar{Y}_{(1)} = 86.6 - 1.6\bar{X}_{(1)}$$

รูปแบบที่ 2: Y สูญหาย 3 คน

$$\bar{Y}_{(2)} = 86.6 - 1.6\bar{X}_{(2)}$$

วิธีการในการจัดการ MNAR

- แทนค่า $\bar{X}$ ในแต่ละรูปแบบเพื่อหา $\bar{Y}$

$$1: \bar{Y}_{(1)} = 86.6 - 1.6(7.25) = 75$$

$$2: \bar{Y}_{(2)} = 86.6 - 1.6(5.6) = 77.64$$

- เอาผลมารวมกันด้วยการเฉลี่ยแบบถ่วงน้ำหนัก

$$\bar{Y} = \frac{4}{7}\bar{Y}_{(1)} + \frac{3}{7}\bar{Y}_{(2)} = \frac{4}{7}(75) + \frac{3}{7}(77.64) = 76.13$$

วิธีการในการจัดการ MNAR

- การแทนค่าข้อมูลที่ประมาณค่าไม่ได้ มีอีกหลายวิธี
- วิธี Pattern mixture model จะทำนายค่าพารามิเตอร์ได้ถูกต้อง เมื่อการแทนค่าทำได้ถูกต้อง ซึ่งเป็นข้อตกลงเบื้องต้นที่ทดสอบไม่ได้

> Rather than rely heavily on poorly estimated MNAR models, I would prefer to examine auxiliary variables that may be related to missingness … and include them in a richer imputation model under assumption of MAR (Schafer, 2003, p. 30)

การออกแบบจงใจให้มีข้อมูลสูญหาย

- การออกแบบจงจงให้มีข้อมูลสูญหาย (Planned missing data designs) คือ การจงใจให้เกิดข้อมูลสูญหาย
  - ลดค่าใช้จ่ายในการเก็บข้อมูล
  - ลดความเหนื่อยล้าในการตอบคำถาม
  - ลดผลของการฝึก (Practice effect)
- แท้จริงแล้ว สถิติที่ใช้กันในปัจจุบัน มีการจงใจให้ข้อมูลสูญหายเช่นกัน

การออกแบบจงใจให้มีข้อมูลสูญหาย

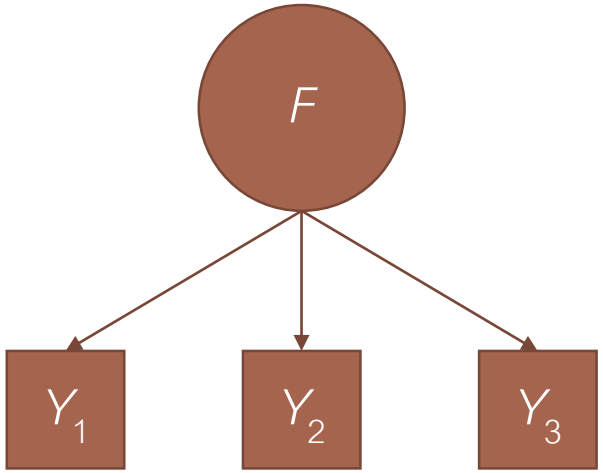
- Independent *t*-test

X	Y with Tx	Y without Tx
T	78	?
T	87	?
T	65	?
T	70	?
C	?	65
C	?	55
C	?	70

การออกแบบจงใจให้มีข้อมูลสูญหาย

- Factor analysis

Y_1	Y_2	Y_3	F
5	4	3	?
5	4	4	?
4	5	4	?
3	5	3	?
3	3	3	?
5	3	4	?
2	4	3	?



การออกแบบจงใจให้มีข้อมูลสูญหาย

- Cohort sequential design

Cohort	2557	2558	2559	2560
2527	43.5	44.9	45.7	46.8
2528	34.9	36.8	40.7	42.2
2529	35.5	40.2	39.8	41.4

Cohort	Age 28	Age 29	Age 30	Age 31	Age 32	Age 33
2527	?	?	43.5	44.9	45.7	46.8
2528	?	34.9	36.8	40.7	42.2	?
2529	35.5	40.2	39.8	41.4	?	?

การออกแบบจงใจให้มีข้อมูลสูญหาย

- การออกแบบจงจงให้มีข้อมูลสูญหาย (Planned missing data designs) เป็นการสุ่มให้ผู้ร่วมการทดลองอยู่ในเงื่อนไข ที่ต้องตอบคำถามชุดหนึ่ง หรือมาเข้าร่วมเฉพาะกลุ่มเวลาที่กำหนด เช่น
  - Three-form design
  - Two-method measurement design

การออกแบบจงใจให้มีข้อมูลสูญหาย

- Three-form design

ชุดที่	ข้อมูลกลุ่ม X	ข้อมูลกลุ่ม A	ข้อมูลกลุ่ม B	ข้อมูลกลุ่ม C
1	วัด	วัด	วัด	ไม่วัด
2	วัด	วัด	ไม่วัด	วัด
3	วัด	ไม่วัด	วัด	วัด

ข้อมูลกลุ่ม X จะเป็นกลุ่มที่ทุกคนตอบ ควรเป็นตัวแปรที่สำคัญในงานวิจัย หรือตัวแปรที่สัมพันธ์กับตัวแปรอื่นๆ สูง

ข้อคำถามจากมาตรเดียวกัน ควรถูกกระจายเข้าสู่กลุ่มต่างๆ (X, A, B, C)

การออกแบบจงใจให้มีข้อมูลสูญหาย

- Three-form design
  - ถึงแม้จะมีชื่อเรียกว่ามี 3 ชุด แต่จะทำให้มีกี่ชุดก็ได้
  - อาจมีชุดหนึ่ง ที่มีผู้ร่วมการทดลองบางส่วน (เช่น 10%) ที่วัดข้อมูลทุกชุด ซึ่งจะทำให้การแทนค่าสูญหายแม่นยำมากขึ้น มีกำลังสูงขึ้นมาก

ชุดที่	ข้อมูลกลุ่ม X	ข้อมูลกลุ่ม A	ข้อมูลกลุ่ม B	ข้อมูลกลุ่ม C	ข้อมูลกลุ่ม D
1	วัด	วัด	วัด	วัด	ไม่วัด
2	วัด	วัด	วัด	ไม่วัด	วัด
3	วัด	วัด	ไม่วัด	วัด	วัด
4	วัด	ไม่วัด	วัด	วัด	วัด
5	วัด	วัด	วัด	วัด	วัด

การออกแบบจงใจให้มีข้อมูลสูญหาย

- Two-method measurement design
  - ใช้เมื่อมีวิธีการวัดที่มีคุณภาพสูง แต่ใช้เวลานานหรือแพง
  - เราอาจใช้วิธีการวัดที่มีคุณภาพต่ำกว่า แต่ถูกหรือใช้เวลาน้อยกว่าเสริมมาตรวัดที่แพง
  - ทำให้เก็บข้อมูลได้มากขึ้นและค่าใช้จ่ายโดยรวมถูกลง เมื่อเทียบกับกำลังทางสถิติที่ได้
  - เช่น ความเครียด อาจวัดโดยใช้ Cortisol และมาตรวัดรายงานตนเอง
  - เช่น เชาวน์ปัญญา อาจวัดโดยใช้ WAIS และมาตรวัดแบบเป็นกลุ่ม

การออกแบบจงใจให้มีข้อมูลสูญหาย

- Two-method measurement design
  - วิธีคือ ให้ผู้ร่วมการทดลองได้รับมาตรวัดที่ถูกทุกคน
  - ให้ผู้ร่วมการทดลองบางคน ได้รับมาตรวัดที่แพง
  - หากใช้วิธีการวิเคราะห์ข้อมูลที่เหมาะสม จะทำให้ Two-method measurement design ได้กำลังสูงขึ้น ขณะที่ค่าใช้เก็บข้อมูลเท่าเดิม

การออกแบบจงใจให้มีข้อมูลสูญหาย

- การออกแบบจงใจให้มีข้อมูลสูญหาย อาจไปประยุกต์ใช้กับการเก็บข้อมูลระยะยาว เช่น การออกแบบการเก็บ Cortisol 7 ครั้ง พบวิธีนี้ดีที่สุด

	วัด baseline	ฝึก Juggling	วัดการตอบสนองต่อความเครียด			ระยะพักฟื้น	
ชุดที่	1 (-20)	2 (0)	3 (+30)	4 (+45)	5 (+60)	6 (+75)	7 (+90)
1	วัด	วัด	วัด	วัด	วัด	วัด	วัด
2	ไม่วัด	วัด	วัด	วัด	ไม่วัด	วัด	วัด
3	วัด	ไม่วัด	วัด	วัด	วัด	ไม่วัด	วัด
4	วัด	วัด	ไม่วัด	วัด	วัด	วัด	ไม่วัด

โปรแกรมจัดการข้อมูลสูญหาย

- SPSS
- SAS
- R: lavaan, semTools, mitools, mice, Amelia
- Mplus

เชิญถามคำถาม หรือเสนอข้อเสนอแนะ

ขอบคุณครับ