การจัดการข้อมูลสูญหาย (Missing Data Handling)

โมเดลพหุระดับ (Multilevel Modeling)

สันทัด พรประเสริฐมานิต

โครงร่างการนำเสนอ

- รูปแบบของค่าสูญหาย
- อคติที่เกิดจากค่าสูญหาย
- วิธีการจัดการค่าสูญหาย
- ตัวอย่างการจัดการค่าสูญหายด้วยการแทนค่าแบบพหุ

รูปแบบของค่าสูญหาย

- ค่าสูญหายเป็นสิ่งที่ต้องพบเจอแทบทุกครั้งในการเก็บข้อมูล โดยผลกระทบของค่าสูญหายแตกต่างกันไปตามลักษณะการเกิดของค่าสูญหาย
- ผลกระทบของค่าสูญหาย อาจอยู่ในรูปของ
  - อคติในการประมาณค่าพารามิเตอร์ (Bias in Parameter Estimates)
  - อคติในการประมาณค่าผิดพลาดมาตรฐานของค่าพารามิเตอร์ (Bias in Standard Error Estimates)
  - สิ่งเหล่านี้ล้วนส่งผลต่อความผิดพลาดแบบที่ 1 (Type I Error) และกำลังในการทดสอบทางสถิติ

รูปแบบของค่าสูญหาย

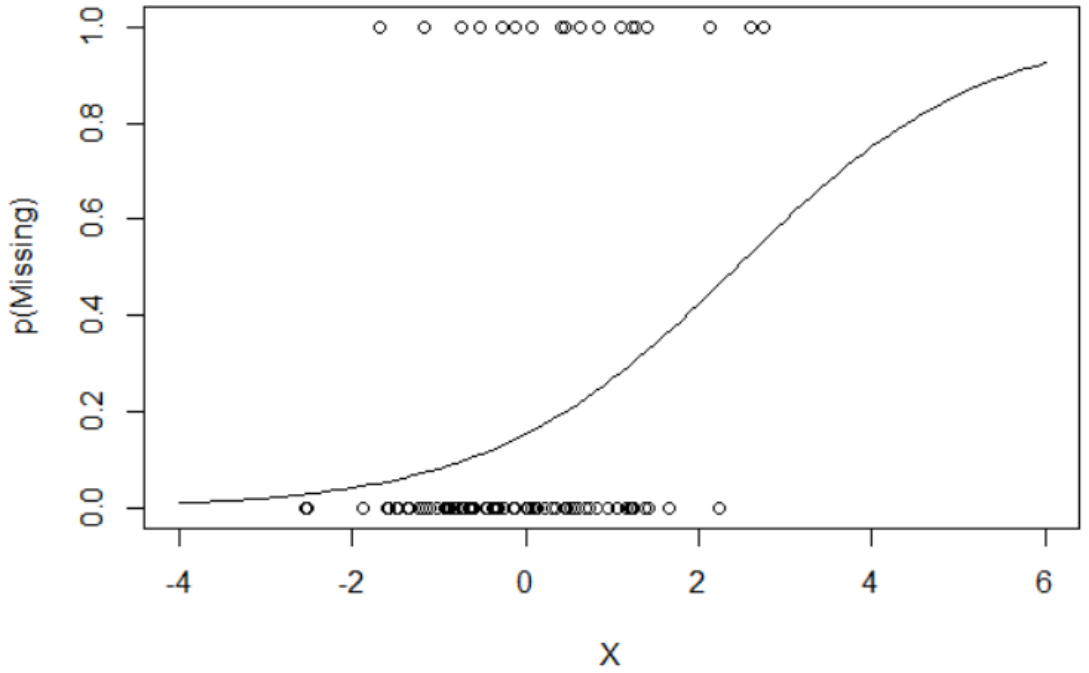
- ให้ M_Y คือ การเกิดค่าสูญหาย โดย 1 = สูญหาย และ 0 = วัดได้ การเกิดค่าสูญหายสามารถแบ่งได้จากแหล่งการเกิดค่าสูญหาย ดังนี้
  - ค่าสูญหายเกิดโดยสุ่มอย่างแท้จริง กล่าวคือ $P(M_Y = 1) = k$ โดย k เป็นค่าคงที่ที่มีค่าตั้งแต่ 0 ถึง 1 เช่น เครื่องคอมเสีย บันทึกข้อมูลตอบไม่ได้
  - ค่าสูญหายเกิดโดยสุ่ม จากอิทธิพลตัวแปรอื่น เช่น เพศหญิง มีแนวโน้มไม่ตอบเรื่องความก้าวร้าวในครอบครัวมากกว่าเพศชาย

$$P(M_Y = 1) = \frac{1}{1 + \exp\left(-(b_0 + b_1 X_1 + \cdots + b_n X_n)\right)}$$

  - ค่าสูญหายเกิดโดยสุ่ม จากอิทธิพลของตนเอง เช่น คนที่มีรายได้สูง จะไม่ตอบรายได้ของตนเอง

$$P(M_Y = 1) = \frac{1}{1 + \exp\left(-(b_0 + b_1 Y)\right)}$$

อิทธิพลของ X ไปหาการเกิดค่าสูญหายของ Y (M_Y)
อยู่ในรูปของการวิเคราะห์ถดถอยแบบลอจิสติค (Logistic Regression)



$$P(M_Y = 1) = \frac{1}{1 + \exp\left(-(-1.7 + 0.7X)\right)}$$

รูปแบบของค่าสูญหาย

- การเกิดค่าสูญหายสามารถแบ่งได้จากวิธีการจัดการ
  - **ค่าสูญหายแบบสุ่มอย่างสมบูรณ์** (Missing Completely at Random; MCAR) เกิดในกรณีที่ค่าสูญหายเกิดจากการสุ่ม ไม่ได้เกี่ยวข้องกับตัวแปรอื่นเลย
    - ตัดข้อมูลของคนที่สูญหายออก นำไปวิเคราะห์ผล ยังไม่มีผลให้เกิดอคติในการประมาณค่าพารามิเตอร์ แต่เกิดอคติในการประมาณค่า *SE* (ซึ่งจะไม่เกิดอคติถ้าจัดการได้ถูกต้อง)
  - **ค่าสูญหายแบบสุ่ม** (Missing at Random; MAR) เกิดในกรณีที่ค่าสูญหายเกิดจากอิทธิพลตัวแปรอื่น ที่วัดได้และจัดการถูกต้อง กล่าวคือ
    - ตัวแปรอื่นที่เก็บข้อมูลมา จะต้องอธิบายกระบวนเกิดค่าสูญหายได้ทั้งหมด
    - นักวิจัยนำข้อมูลตัวแปรอื่นที่อธิบายกระบวนการเกิดค่าสูญหาย มาใช้ในการจัดการค่าสูญหายได้ ด้วยวิธีการที่ถูกต้อง จะไม่เกิดอคติทั้งการประมาณค่าพารามิเตอร์ และ *SE*

รูปแบบของค่าสูญหาย

- การเกิดค่าสูญหายสามารถแบ่งได้จากวิธีการจัดการ
  - **ค่าสูญหายแบบไม่สุ่ม** (Missing not at Random; MNAR) เกิดขึ้นในกรณีที่ค่าสูญหายเกิดจากอิทธิพลตัวแปรอื่นหรือค่าของตนเอง อาจเกิดขึ้นในกรณีดังต่อไปนี้
    - เก็บข้อมูลตัวแปรที่มีผลต่อค่าสูญหายมา แต่ไม่ใช้วิธีการจัดการค่าสูญหายที่ถูกต้อง
    - เก็บข้อมูลตัวแปรที่มีผลต่อค่าสูญหายมาแค่บางส่วน แต่ไม่ได้อธิบายกระบวนการเกิดค่าสูญหายทั้งหมด แม้จะจัดการได้ถูกต้องแล้วก็ตาม ก็ยังเป็น MNAR
    - ไม่ได้เก็บข้อมูลตัวแปรที่เกี่ยวข้องกับกระบวนการเกิดค่าสูญหายมา
    - การเกิดค่าสูญหาย เกิดจากค่าของตัวแปรดังกล่าวเอง ทำให้ไม่มีตัวแปรในการอธิบายกระบวนการเกิดค่าสูญหาย
    - จะเกิดอคติต่อทั้งการประมาณค่าพารามิเตอร์และ *SE*

รูปแบบของค่าสูญหาย

- ถ้าจัดการด้วยวิธีการที่ถูกต้อง ค่าสูญหายที่เป็นแบบ MCAR และ MAR จะไม่เกิดอคติในการประมาณค่าพารามิเตอร์และ *SE* แต่ถ้าเป็น MNAR จะเกิดอคติ
- นักวิจัย**ไม่มีทาง**ที่จะทราบว่าข้อมูลของตนเองเป็น MAR หรือ MNAR เพราะนักวิจัยไม่รู้ว่ากระบวนการเกิดค่าสูญหาย เกิดได้อย่างไร
- ในความเป็นจริงแล้ว ข้อมูลแทบทั้งหมด ไม่ได้เป็น MAR 100% เพียงแค่มีระดับว่าค่อนไปทาง MAR มากน้อยเพียงใด
- ข้อมูลหนึ่งจะมีหลายตัวแปร ซึ่งระดับความเป็น MAR ย่อมต่างกันระหว่างตัวแปร

รูปแบบของค่าสูญหาย

- ด้วยเหตุนี้ การจัดการข้อมูลสูญหาย ไม่จำเป็นมาคัดแยกว่าข้อมูลนี้เป็น MCAR, MAR, หรือ MNAR เพราะคุณไม่มีทางรู้อยู่แล้ว
- สิ่งที่ทำได้ คือ
  - เก็บข้อมูลที่น่าจะทำนายการเกิดข้อมูลสูญหายมาให้ได้มากที่สุด เพื่อให้ข้อมูลที่เก็บได้เป็น MAR มากที่สุด
  - ใช้วิธีการที่ดีที่สุด ในการจัดการข้อมูลสูญหาย เพื่อลดอคติจากค่าสูญหาย
  - ถ้าการเกิดข้อมูลสูญหาย เกิดจากตัวแปรดังกล่าวเอง ควรหาตัวแปรอื่นเสริม ที่น่าจะทำนายข้อมูลสูญหายช่องดังกล่าว ให้ดีที่สุด เช่น ไม่มีข้อมูลน้ำหนัก ก็หาข้อมูลความสูง การกินของหวาน รอบเอว ฯลฯ มาช่วยทำนายค่าสูญหายที่หายไป และลดอคติจากการขาดค่าสูญหายไปได้

อคติที่เกิดจากค่าสูญหาย

- เพื่อให้เห็นภาพ สมมติว่ามีข้อมูลที่สร้างจากกระบวนการดังต่อไปนี้

$$Y_{ij} = 50 + 5\left(X_{ij} - \bar{X}_{..}\right) + u_{0j} + e_{ij}$$
$$u_{0j} \sim N(0,7) \quad e_{ij} \sim N(0,7)$$

  - โดย $X_{ij} = 0.5 + v_{0j} + r_{ij}$, $v_{0j} \sim N(0, 0.7)$, $r_{ij} \sim N(0,1)$
- ต้องการทำนายค่า γ_{00} ในโมเดลนี้

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

ตัวอย่างอคติที่เกิดจากค่าสูญหาย

- ข้อมูลถูกสร้างขึ้นมา โดยมีจำนวนกลุ่ม 1,000 กลุ่ม และมีคนกลุ่มละ 100 คน
- จากข้อมูล 100,000 คน ถูกทำให้เกิดค่าสูญหายไป 30% ของตัวแปร Y ใน 3 สถานการณ์ดังนี้
  - สถานการณ์ที่ 1: $P(M_Y = 1) = .3$
  - สถานการณ์ที่ 2: $P(M_Y = 1) = \frac{1}{1+\exp(-(-1.015+X))}$
  - สถานการณ์ที่ 3: $P(M_Y = 1) = \frac{1}{1+\exp(-(-1.015+Y))}$

ตัวอย่างอคติที่เกิดจากค่าสูญหาย

สถานการณ์ที่ 1 เป็นสถานการณ์ MCAR ซึ่งอคติในการประมาณค่า γ_{00} และ $SE(\gamma_{00})$ ไม่เกิดขึ้น

สถานการณ์ที่ 2 เป็นสถานการณ์ MNAR เพราะ X ทำนายการสูญหายของ Y แต่ไม่มี X ในโมเดล

โมเดล	γ_{00}	$SE(\gamma_{00})$
ไม่มีค่าสูญหาย (จุดอ้างอิง)	52.490	0.257
สถานการณ์ที่ 1 M_Y เกิดโดยสุ่ม	52.485	0.258
สถานการณ์ที่ 2 M_Y ได้รับอิทธิพลจาก X	51.359	0.253
สถานการณ์ที่ 3 M_Y ได้รับอิทธิพลจาก Y	50.778	0.236

ใช้โมเดลว่างทำนายค่าเฉลี่ย

สถานการณ์ที่ 3 เป็นสถานการณ์ MNAR เพราะ Y ทำนายการสูญหายของ Y ซึ่งค่า Y ที่สูญหายไปไม่สามารถเก็บข้อมูลได้

ตัวอย่างอคติที่เกิดจากค่าสูญหาย

โมเดล	γ_{00}	$SE(\gamma_{00})$
ไม่มีค่าสูญหาย (จุดอ้างอิง)	52.490	0.257
สถานการณ์ที่ 1 M_Y เกิดโดยสุ่ม	52.485	0.258
สถานการณ์ที่ 2 M_Y ได้รับอิทธิพลจาก X	51.359	0.253
สถานการณ์ที่ 3 M_Y ได้รับอิทธิพลจาก Y	50.778	0.236
สถานการณ์ที่ 2 ใส่ตัวแปร X ทำนาย (ย้ายศูนย์กลางที่ค่าเฉลี่ยรวม)	52.463	0.231

สังเกตว่า สถานการณ์ที่ 2 และ 3 เกิดอคติในการประมาณค่า เนื่องจากเกิดสถานการณ์ MNAR ขึ้น

สถานการณ์ในแถวล่างสุด เป็น MAR เนื่องจากมีการใส่ X เข้าไปในโมเดล แต่อย่างไรก็ตามก็ไม่ใช่การจัดการค่าสูญหายที่ถูกต้อง เพราะ γ_{00} ในจุดอ้างอิง คือ ค่าเฉลี่ยรวม แต่ γ_{00} ในแถวล่างสุด คือ ค่าเฉลี่ยของ Y เมื่อ X เท่ากับค่าเฉลี่ยรวม แม้จะใกล้เคียงกัน แต่เป็นคนละสิ่งกัน สังเกตได้ว่า SE ไม่เท่ากับจุดอ้างอิง

วิธีการจัดการค่าสูญหาย

- ในสไลด์ที่ผ่านมา การจัดการค่าสูญหายที่จัดทำไป คือ การลบทั้งแถว (Listwise Deletion) ซึ่งจะจัดการค่าสูญหายได้ดี ในกรณีเดียว คือ MCAR
- วิธีการที่จัดการ MAR ได้ มี 2 วิธี คือ
  - วิธีการคำนวณความเป็นไปได้สูงสุดโดยตรง (Direct Maximum Likelihood; ML)
  - วิธีการแทนค่าเป็นหลายข้อมูล (Multiple Imputation; MI)
- ในการวิเคราะห์พหุระดับแบบปกติ วิธีการที่เหมาะสมที่สุด คือ MI เนื่องจากการวิเคราะห์พหุระดับ จะวิเคราะห์หลายโมเดล และแต่ละโมเดลมีจำนวนตัวแปรที่แตกต่างกัน
- ML ทำได้ใน MSEM เหมาะสำหรับโมเดลที่วิเคราะห์ด้วย MSEM ตั้งแต่ต้น

วิธีการจัดการค่าสูญหาย

- วิธีการแทนค่าแบบพหุ (Multiple Imputation; MI) คือ การแทนค่าข้อมูลสูญหายรูปแบบหนึ่ง ซึ่งจะแทนค่าข้อมูลสูญหาย ด้วยค่าที่ทำนายได้บวกกับความผิดพลาดในการทำนาย
- เพื่อไม่ให้ผลการวิเคราะห์ได้รับอิทธิพลจากความผิดพลาดในการทำนายที่สุ่มออกมาเพียงแค่ตัวเดียว นักวิจัยจึงแทนค่าออกมาหลายครั้ง ได้ข้อมูลหลายชุด
- นำข้อมูลแต่ละชุดมาวิเคราะห์สถิติที่ต้องการ ได้ค่าสถิติที่ต้องการจากข้อมูลหลายชุด
- นำค่าสถิติที่ได้ทั้งหมด มารวมกัน และหา *SE* ที่ถูกต้อง ด้วยวิธีการของ Rubin

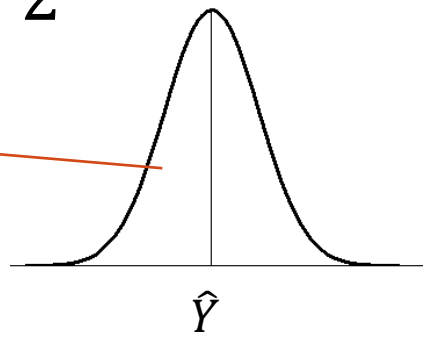
X	Y	Z
5	7	5
4	8	4
3	?	1
4	7	2
6	9	2

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

สมมติว่า $SE(\hat{Y}) = 2$

สุ่มค่าจากโค้งปกติ

จาก $N(7.17, 2)$



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% การใช้ Listwise Deletion ไม่ได้มีอคติมาก แต่ถ้าใช้ MI ได้ ให้ใช้ดีกว่า
- แต่หากตัวแปรใดตัวแปรหนึ่ง มีข้อมูลสูญหายเกิน 5% ควรใช้ MI
- ในข้อมูลแบบพหุระดับ ให้ใช้ `mice` package ใน R จัดการข้อมูลสูญหายได้ด้วยวิธี MI
- MI จะทำให้ผู้วิจัยได้ข้อมูลกลุ่มหนึ่ง (มีข้อมูลที่ถูกแทนค่าแล้วหลายข้อมูล) เพื่อสามารถนำไปวิเคราะห์ข้อมูล หาค่าสถิติ แล้วรวมค่าสถิติในภายหลัง

วิธีการใหม่การจัดการข้อมูลสูญหาย

- อย่างไรก็ตาม วิธีการทำ MI มีหลายวิธี วิธีการที่ผมจะพูดถึงในงานครั้งนี้ คือ วิธีการกำหนดค่าแบบเงื่อนไขสมบูรณ์ (Fully Conditional Specification; FCS)
- วิธีการดั้งเดิม คือ จะต้องหาการกระจายที่เหมาะสมกับข้อมูลทั้งหมดทุกตัวแปรให้ได้ก่อน (เช่น การกระจายทั้งหมดเป็น Multivariate Normal Distribution) แล้วค่อยสุ่มข้อมูลลงจุดที่สูญหาย วิธีนี้เรียกว่าวิธีโมเดลการกระจายร่วม (Joint Modelling) ซึ่งยังไม่สามารถจัดการความชันสุ่มได้

วิธีการใหม่การจัดการข้อมูลสูญหาย

- วิธี FCS จะหาสมการทำนายค่าสูญหายตัวแปรแรก จากตัวแปรอื่นๆ แล้ววนไปทำนายตัวแปรที่สอง (อิงตัวแปรแรกที่แทนค่าสูญหายแล้ว) ไปยังตัวแปรที่สาม ไปจนตัวแปรสุดท้าย แล้ววนกลับตัวแปรที่ 1 ไปเรื่อยๆ โดยผู้วิเคราะห์จะกำหนดจำนวนชุดข้อมูล และจำนวนรอบในการวนได้
- วิธี FCS เป็นวิธีการที่เข้าใจง่าย คำนวณได้รวดเร็ว จัดการข้อมูลขนาดใหญ่ได้ แต่วิธีการนี้ยังไม่มีทฤษฎีทางสถิติอย่างสมบูรณ์มาสนับสนุน แต่จากการทดสอบการโมเดลจำลอง (Simulation) แล้ว พบว่าวิธี FCS สามารถประมาณค่าสูญหายได้ดี

เริ่มต้น

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

สุ่มค่าเริ่มต้น

X	Y	Z
5	2.9	5
4	8	2.5
3.5	3.4	1
4.7	7	2
6	9	7.1

หา $\hat{X} + e$

X	Y	Z
5	2.9	5
4	8	2.5
3.7	3.4	1
4.2	7	2
6	9	7.1

หา $\hat{Y} + e$

X	Y	Z
5	3.2	5
4	8	2.5
3.7	4.1	1
4.2	7	2
6	9	7.1

หา $\hat{Z} + e$

X	Y	Z
5	3.2	5
4	8	2.9
3.7	4.1	1
4.2	7	2
6	9	6.5

หา $\hat{X} + e$

X	Y	Z
5	3.2	5
4	8	2.9
3.2	4.1	1
4.1	7	2
6	9	6.5

หา $\hat{Y} + e$

X	Y	Z
5	3.5	5
4	8	2.9
3.2	4.5	1
4.1	7	2
6	9	6.5

หา $\hat{Z} + e$

X	Y	Z
5	3.5	5
4	8	3.0
3.2	4.5	1
4.1	7	2
6	9	6.8

แทนค่าวนไปเรื่อยๆ หลายๆ รอบ (เช่น 20 รอบ)

ได้ผลการแทนค่าชุดที่ 1

X	Y	Z
5	3.6	5
4	8	2.9
3.1	4.7	1
3.9	7	2
6	9	6.5

การแทนค่าชุดที่ 1

X	Y	Z
5	3.6	5
4	8	2.9
3.1	4.7	1
3.9	7	2
6	9	6.5

การแทนค่าชุดที่ 2

X	Y	Z
5	3.7	5
4	8	2.5
3.5	4.3	1
3.9	7	2
6	9	6.4

การแทนค่าชุดที่ 3

X	Y	Z
5	3.5	5
4	8	3.2
3.7	4.0	1
4.0	7	2
6	9	6.3

การแทนค่าชุดที่ 4

X	Y	Z
5	3.6	5
4	8	2.8
3.4	4.1	1
3.9	7	2
6	9	6.5

การแทนค่าชุดที่ 5

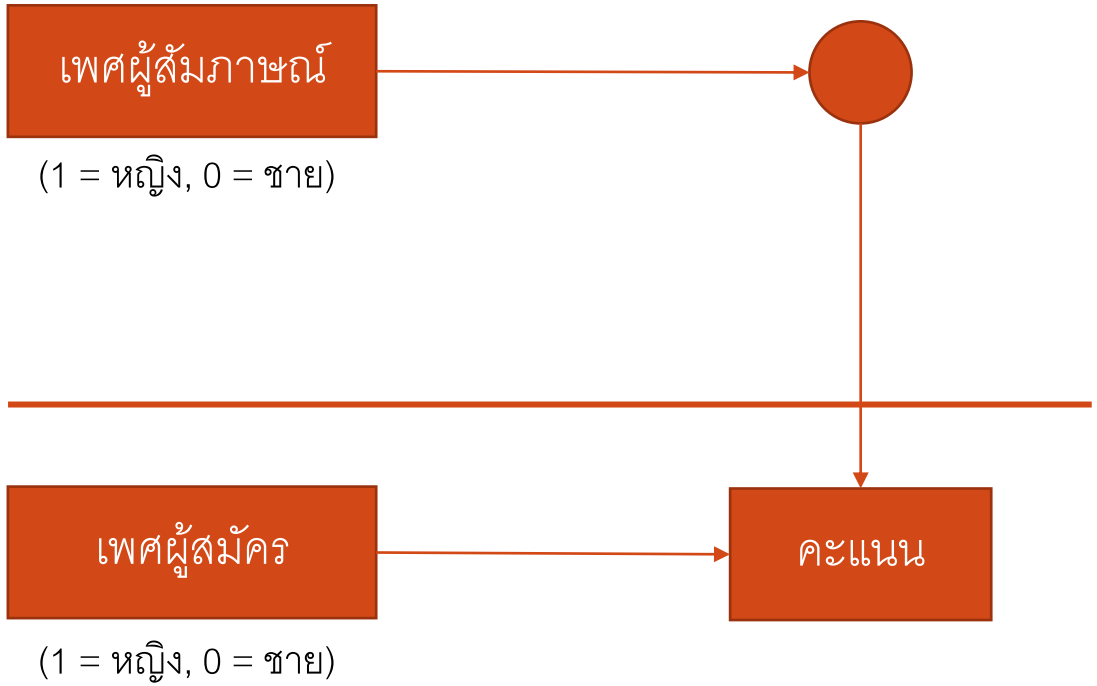
X	Y	Z
5	3.4	5
4	8	2.9
3.8	4.5	1
4.0	7	2
6	9	6.3

- การแทนค่าแบบพหุ จะแทนค่าสูญหายออกมาหลายๆ ชุด แต่ละชุดเกิดจากค่าเริ่มต้นที่แตกต่างกัน และการสุ่มค่าความผิดพลาด (error; *e*) ที่แถมไปกับการแทนค่าแต่ละครั้งที่แตกต่างกัน
- ค่าสูญหายบางตำแหน่ง จะแกว่งไปแกว่งมาเยอะมาก แสดงว่าความแม่นยำในการทำนายข้อมูลตำแหน่งดังกล่าวไม่แม่น แต่ค่าสูญหายบางตำแหน่งก็แกว่งไปแกว่งมาน้อย แสดงว่าการแทนค่าดังกล่าวค่อนข้างแม่น
- ความแม่นของข้อมูลแต่ละจุด จะสะท้อนออกมาที่ผลการวิเคราะห์ข้อมูลที่สนใจ ว่าค่าพารามิเตอร์ที่ทำนายได้ (Parameter Estimates) และค่าความผิดพลาดมาตรฐาน (Standard Error) แตกต่างกันระหว่างชุดข้อมูลที่แทนค่าหรือไม่
- Rubin's Rule ที่ใช้รวมผลการวิเคราะห์ จะได้ผลค่าความผิดพลาดมาตรฐานรวม (Pooled Standard Errors) สูง ถ้าค่าพารามิเตอร์ที่ทำนายได้แกว่งไปแกว่งมามาก แต่ถ้าค่าพารามิเตอร์ที่ทำนายได้แกว่งไปมาน้อยมาก ค่าความผิดพลาดมาตรฐานรวม ก็จะใกล้เคียงค่าความผิดพลาดมาตรฐานของแต่ละชุด

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

ทดสอบอคติของผู้สัมภาษณ์
ว่าผู้สัมภาษณ์ประเมินผู้สมัคร
เพศเดียวกับผู้สัมภาษณ์สูงขึ้นหรือไม่

ตัวอย่างที่ 1 มีค่าสูญหายเพียงตัวแปรเดียว



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

วิเคราะห์รูปแบบค่าสูญหาย (Missing Data Pattern)

```
> dat1 <-read.table("lecture11ex1.csv", sep=",", header=TRUE, na.strings="999999")
> dat1a <- dat1[,c("erid", "score", "eesex", "ersex")]
> library(mice)
> md.pattern(dat1a)
     erid eesex ersex score
7907    1     1     1     1    0
2093    1     1     1     0    1
        0     0     0  2093 2093
```

คัดแค่ 4 ตัวแปรเป็นตัวอย่าง

มี 7,907 คนที่มีข้อมูลครบทุกค่า

มี 2,093 คนที่ข้อมูลคะแนนสูญหายไป

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนหลักมากๆ ให้ย้ายศูนย์กลางของตัวแปรต่อเนื่องมาที่ค่าเฉลี่ย (ทำให้ค่าเฉลี่ยของตัวแปรใหม่เท่ากับ 0)

```
> m_score1a <- mean(dat1a$score, na.rm=TRUE)
> dat1a$score <- dat1a$score - m_score1a
```

- ค่าของตัวแปรระดับที่ 2 ไม่สามารถคำนวณค่าเฉลี่ยโดยการใช้คำสั่ง `mean` ตรงๆ ได้ (ดูตัวอย่างถัดไป)
- ค่าเฉลี่ยที่หักออกไป จะบวกกลับในภายหลัง

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> pred1a <- make.predictorMatrix(dat1a)
> pred1a
       erid score eesex ersex
erid      0     1     1     1
score     1     0     1     1
eesex     1     1     0     1
ersex     1     1     1     0
```

คอลัมน์แทนตัวแปรทำนาย

แถวแทนตัวแปรที่ถูกแทนค่าสูญหาย

```
> meth1a <- make.method(dat1a)
> meth1a
 erid score eesex ersex
   "" "pmm"    ""    ""
```

สถิติที่ใช้ในการแทนค่าสูญหายในแต่ละตัวแปร

```
> pred1a <- make.predictorMatrix(dat1a)
> pred1a
       erid score eesex ersex
erid      0     1     1     1
score     1     0     1     1
eesex     1     1     0     1
ersex     1     1     1     0
```

ตัวเลขในเมทริกซ์มีความหมายดังนี้

0 = ไม่ต้องใช้ตัวแปรคอลัมน์ในการทำนายตัวแปรแถว

-2 = ใช้ตัวแปรคอลัมน์เป็น ID ของระดับที่ 2 ในการทำนายตัวแปรแถว

1 = ตัวแปรคอลัมน์ทำนายตัวแปรแถวตรงๆ ไม่มีความชันสุ่ม ไม่ใส่ค่าเฉลี่ยกลุ่ม

2 = ตัวแปรคอลัมน์ทำนายตัวแปรแถว โดยมีความชันสุ่ม แต่ไม่ใส่ค่าเฉลี่ยกลุ่ม

3 = ตัวแปรคอลัมน์ทำนายตัวแปรแถว ไม่มีความชันสุ่ม ใส่ค่าเฉลี่ยกลุ่ม

4 = ตัวแปรคอลัมน์ทำนายตัวแปรแถว มีความชันสุ่ม ใส่ค่าเฉลี่ยกลุ่ม

กรณีคอลัมน์เป็นตัวแปรระดับที่ 1

$$1:\ Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + e_{ij}$$

$$2:\ Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

$$3:\ Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}\bar{X}_{.j} + u_{0j} + e_{ij}$$

$$4:\ Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}\bar{X}_{.j} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

กรณีคอลัมน์เป็นตัวแปรระดับที่ 2

$$1:\ Y_{ij} = \gamma_{00} + \gamma_{01}W_{.j} + u_{0j} + e_{ij}$$

การกำหนด เลข 2, 3, 4 ทำไม่ได้ เพราะเป็นตัวแปรระดับที่ 2 ไม่มีค่าเฉลี่ยกลุ่มและความชันสุ่ม

```
> pred1a <- make.predictorMatrix(dat1a)
> pred1a
       erid score eesex ersex
erid      0     1     1     1
score
eesex     1     1     0     1
ersex     1     1     1     0
```

แนวทางในการใส่ตัวเลข 0 เป็นดังนี้

1. ตัวแปรที่ไม่มีค่าสูญหาย ทุกค่าในแถวเท่ากับ 0
2. ตัวแปรที่ไม่ได้ใช้ในการวิเคราะห์ และไม่เกี่ยวข้องกับค่าสูญหาย (ไม่ได้ใช้ทำนายค่าสูญหาย เช่น ตัวแปร ID ในระดับที่ 1 ให้ทุกค่าในคอลัมน์และแถวเท่ากับ 0

```
> pred1a <- make.predictorMatrix(dat1a)
> pred1a
       erid score eesex ersex
erid      0     1     1     1
score    -2     0     4     1
eesex     1     1     0     1
ersex     1     1     1     0
```

แนวทางในการใส่ตัวเลขในแถวที่ตัวแปรมีค่าสูญหายเป็นดังนี้

1. ใส่ -2 สำหรับ ID ในระดับที่ 2
2. ใส่ 0 สำหรับตัวแปรตนเอง เพราะตนเองไม่สามารถทำนายค่าสูญหายตนเองได้
3. ใส่ 0 สำหรับตัวแปรที่ไม่เกี่ยวข้อง
4. ใส่ 1 สำหรับตัวแปรคอลัมน์ทำนายค่าสูญหาย ที่เป็นตัวแปรระดับที่ 2
5. ใส่ 3 หรือ 4 สำหรับตัวแปรคอลัมน์ทำนายค่าสูญหาย ที่เป็นตัวแปรระดับที่ 1

เนื่องจาก `ersex` น่าจะมีความชันสุ่มกับคะแนนการสัมภาษณ์ เลยใช้ 4 ซึ่งเป็นโมเดลที่มีความชันสุ่ม และใส่ค่าเฉลี่ยกลุ่ม แต่จะเห็นว่าการใส่ 4 นี้ คือตัวแปรไม่ได้ทำ Group mean centering ซึ่งการทำ Group mean centering จะกล่าวถึงในตัวอย่างถัดไป

```
> meth1a <- make.method(dat1a)
> meth1a
 erid score eesex ersex
   "" "2l.pan"    ""     ""
```

สถิติที่ใช้ในการทำนายค่าสูญหาย ในโมเดลพหุระดับมีดังนี้

`""` = ไม่ใช้ทำนาย ใช้สำหรับตัวแปรที่ไม่มีค่าสูญหาย

`"2l.pan"` = DV เป็นตัวแปรต่อเนื่อง ระดับที่ 1 ความแปรปรวนแต่ละกลุ่มเท่ากัน

`"2l.norm"` = DV เป็นตัวแปรต่อเนื่อง ระดับที่ 1 ความแปรปรวนแต่ละกลุ่มไม่เท่ากัน

`"2l.bin"` = DV เป็นตัวแปรดัมมี่ ระดับที่ 1

`"2lonly.norm"` = DV เป็นตัวแปรต่อเนื่อง ระดับที่ 2

`"2lonly.pmm"` = DV เป็นตัวแปรดัมมี่ (หรือต่อเนื่อง) ระดับที่ 2

`"2l.groupmean"` = ให้ DV เป็นผลการคำนวณค่าเฉลี่ยกลุ่มของตัวแปร
คอลัมน์ที่กำหนดเป็น 2 และกลุ่มกำหนดด้วยคอลัมน์ที่เป็น -2

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> pred1a[,] <- 0
> pred1a["score",] <- c(-2, 0, 4, 1)
> meth1a["score"] <- "2l.pan"
```

ใส่ script เพื่อเปลี่ยนค่าดังที่กล่าวไปข้างต้น

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. ดำเนินการแทนค่าสูญหาย

สถิติทำนายค่าสูญหาย จำนวนชุดข้อมูลที่แทนค่า ชุดเลขสุ่ม

```
> imp <- mice(dat1a, pred=pred1a, meth=meth1a, m=10, maxit=10, seed=123321)
```

ข้อมูล เมทริกซ์กำหนดตัวแปรทำนาย จำนวนรอบหมุนแทนค่า

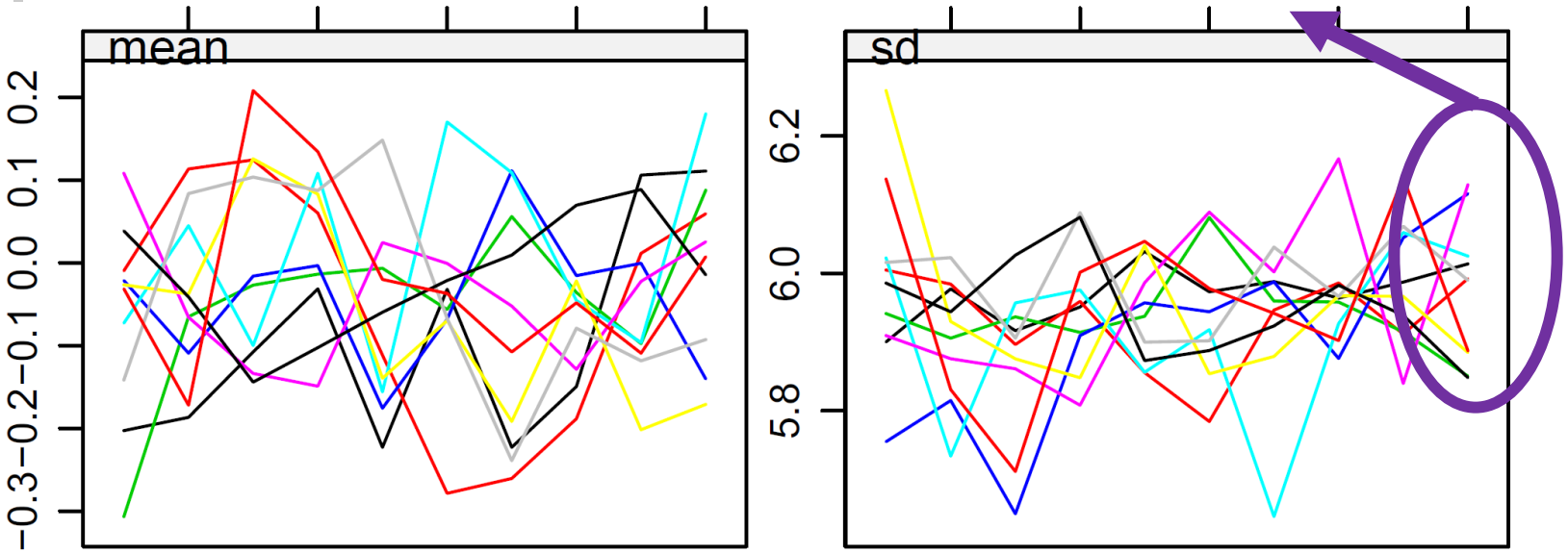
```
 iter imp variable
  1   1  score
  1   2  score
  1   3  score
  1   4  score
  1   5  score
  1   6  score
  1   7  score
  1   8  score
  1   9  score
  1   10  score
  2   1  score
  2   2  score
  2   3  score
  2   4  score
  2   5  score
```

เมื่อสั่งคำสั่งแล้ว โปรแกรมจะแทนค่าสูญหายข้อมูลทีละชุด (imp) วนตามจำนวนรอบ (iter) ที่กำหนด

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. ตรวจสอบการลู่เข้าของการแทนค่า

```
> plot(imp, c("score"))
```



ดูเหมือนจะลู่เข้าในโซนเดียวกัน แต่ไม่ชัดเจน
ตอนปลายเหมือนกระจายออกเล็กน้อยด้วยซ้ำ

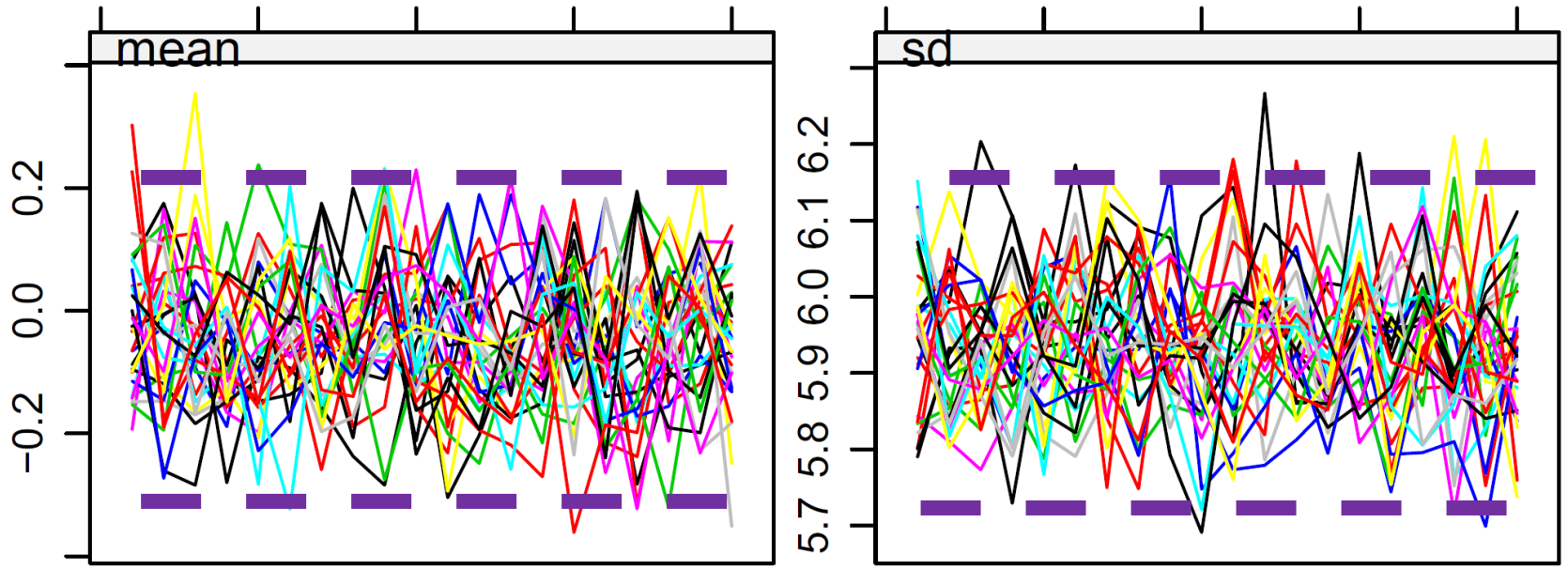
ลู่เข้าเมื่อโซนด้านขวาของกราฟ เส้นกระจัดกระจายอยู่ในโซนเดียวกัน
ทั้งค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. ตรวจสอบการลู่เข้าของการแทนค่า

เพิ่มเป็น 20 รอบการหมุน ข้อมูล 30 ชุด

```
imp <- mice(dat1a, pred=pred1a, meth=meth1a, m=30, maxit=20, seed=123321)
plot(imp, c("score"))
```



ลองลากเส้นขนานกลุ่มเส้น แล้วพบว่าการเปลี่ยนแปลงอยู่ในโซนนั้น ถือว่าลู่เข้า แต่ถ้ามันขยายออกหรือหุบลง แสดงว่าข้อมูลยังไม่ลู่เข้า

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

5. บวกค่าเฉลี่ยที่หักออกมา เข้าไปที่ตัวแปรที่แทนค่าแล้วทั้งหมด

- ผลการวิเคราะห์จะออกมาเป็นวัตถุหนึ่ง ไม่ใช้ข้อมูลดิบที่แทนค่าแล้วโดยตรง
- คำสั่ง `complete` เป็นการสร้างข้อมูลดิบที่แทนค่าแล้ว โดย `"long"` เป็นการสร้างข้อมูลดิบที่แทนค่าแล้ว โดยนำมาซ้อนกันเป็นข้อมูลชุดเดียวกันหลายๆ แถว

```
> impdat1a <- complete(imp, "long", include = TRUE)
> impdat1a$score <- impdat1a$score + m_score1a
> imp1a <- as.mids(impdat1a)
```

บวกค่าเฉลี่ยกลับ

แปลงกลับเป็นผลลัพธ์ที่เกิดจาก `mice`

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

6. วิเคราะห์ข้อมูลที่แทนค่าแต่ละข้อมูล

```
> fit1 <- with(imp1a, lmer(score ~ eesex + ersex + (1|erid), REML=FALSE))
> fit2 <- with(imp1a, lmer(score ~ eesex + ersex + (1 + eesex|erid), REML=FALSE))
```

คำสั่งการวิเคราะห์เหมือนเดิม เพียงแค่ไม่ใส่ `data` ลงไป

เพราะคำสั่ง `with` จะเป็นการกระจาย `data` ไปใช้คำสั่งวิเคราะห์ทั้งหมด

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. รวมผลการวิเคราะห์ข้อมูล

หา Fixed Effect

```
> out1 <- pool(fit1)
> summary(out1)
                estimate  std.error   statistic         df   p.value
(Intercept) 72.74405870 0.20008516 363.5654790 5683.1515 0.0000000
eesex       -0.04686866 0.09668749  -0.4847438  275.7164 0.6282429
ersex        0.44886412 0.27415965   1.6372363 8271.7925 0.1016191
> out2 <- pool(fit2)
> summary(out2)
                estimate std.error   statistic         df   p.value
(Intercept) 72.78282873 0.1987133 366.2705825 4882.0592 0.0000000
eesex       -0.04686866 0.1036825  -0.4520404  361.1187 0.6515113
ersex        0.37132405 0.2739091   1.3556469 7479.7870 0.1752525
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

ให้คัดลอก mimlmtools.R ไปใส่ในโฟลเดอร์เดียวกับไฟล์ข้อมูล

7. รวมผลการวิเคราะห์ข้อมูล

หา Random Effect

ranefMI, sigmaMI, anovaMI เป็นฟังก์ชั่นที่ผมเขียนขึ้นมาเอง ซึ่งสามารถใช้ได้จากการโหลดไฟล์ mimlmtools.R ดังกล่าว

```
> source("mimlmtools.R")
> ranefMI(fit1)
             (Intercept)
(Intercept)    16.62006
attr(,"stddev")
(Intercept)
   4.076771
attr(,"correlation")
             (Intercept)
(Intercept)            1
> ranefMI(fit2)
             (Intercept)     eesex
(Intercept)  16.0975348 0.1737632
eesex         0.1737632 1.5767844
attr(,"stddev")
(Intercept)       eesex
   4.012173    1.255701
attr(,"correlation")
             (Intercept)      eesex
(Intercept)  1.00000000 0.03448989
eesex        0.03448989 1.00000000
> sigmaMI(fit1)
[1] 19.2201
> sigmaMI(fit2)
[1] 18.7821
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. รวมผลการวิเคราะห์ข้อมูล

เปรียบเทียบสองโมเดล

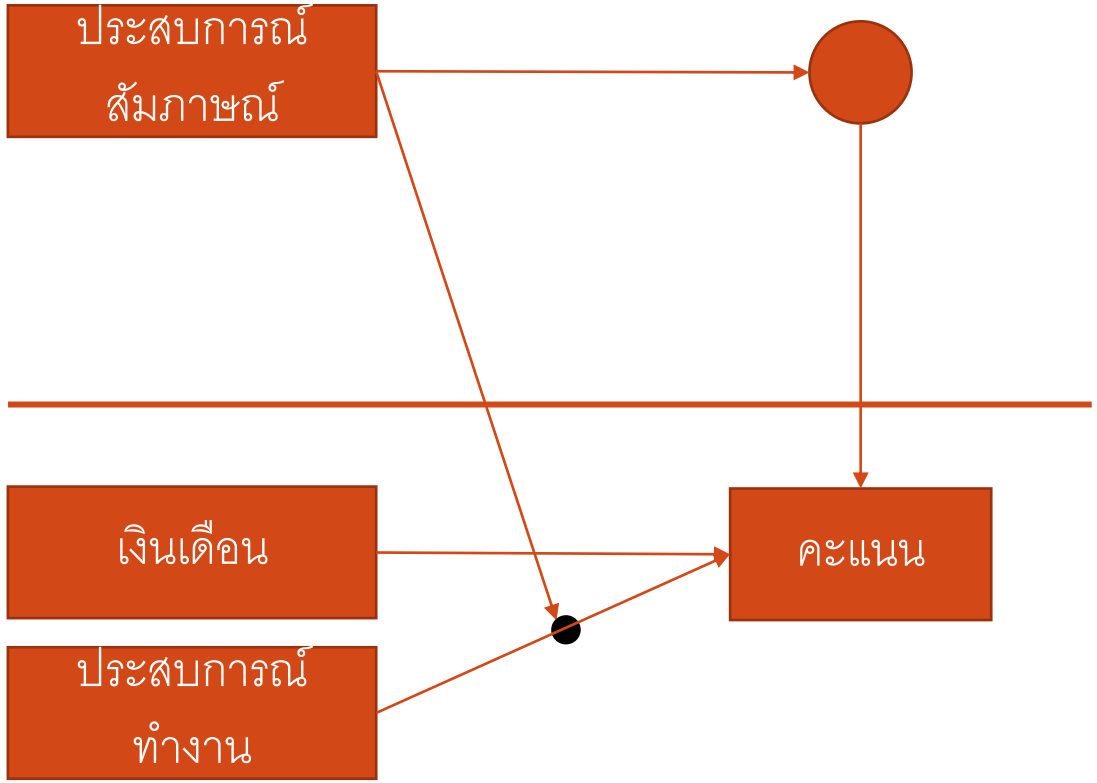
```
> anovaMI(fit1, fit2)
      F     df1      df2     p.F
  4.287   2.000  159.890   0.015
```

วิธีการเปรียบเทียบนี้ จะเรียกว่าสถิติ D2 ตามแนวคิดของ Li, Meng, Raghunathan, & Rubin (1991) ซึ่งเป็นแนวทางในการรวมค่า χ^2 จากข้อมูลหลายๆ ชุด ไว้เป็นค่าสถิติเดียว

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

ทดสอบอคติของผู้สัมภาษณ์
ว่าผู้สัมภาษณ์ประเมินผู้สมัคร
เพศเดียวกับผู้สัมภาษณ์สูงขึ้นหรือไม่

ตัวอย่างที่ 2 มีค่าสูญหายเพียง 4 ตัวแปร
และมีปฏิสัมพันธ์



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

วิเคราะห์รูปแบบค่าสูญหาย (Missing Data Pattern)

```
> dat1b <- dat1[,c("erid", "score", "eesalary", "eeworkexp", "erexp")]
> library(mice)
> md.pattern(dat1b)
     erid eeworkexp eesalary erexp score
4030    1         1        1     1     1    0
1083    1         1        1     1     0    1
1049    1         1        1     0     1    1
257     1         1        1     0     0    2
1007    1         1        0     1     1    1
266     1         1        0     1     0    2
271     1         1        0     0     1    2
60      1         1        0     0     0    3
954     1         0        1     1     1    1
272     1         0        1     1     0    2
278     1         0        1     0     1    2
67      1         0        1     0     0    3
248     1         0        0     1     1    2
70      1         0        0     1     0    3
70      1         0        0     0     1    3
18      1         0        0     0     0    4
        0      1977     2010  2070  2093 8150
```

มี 4,030 คนที่มีข้อมูลครบทุกค่า

ที่เหลือมีค่าสูญหายอย่างน้อย 1 ค่า

มีรูปแบบค่าสูญหายทั้งหมด 15 รูปแบบ

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนมากๆ ให้เปลี่ยนเป็นคะแนนมาตรฐานทั้ง 4 ตัวแปร

หาค่าเฉลี่ย

```
> m_score1b <- mean(dat1b$score, na.rm=TRUE)
> m_eesalary1b <- mean(dat1b$eesalary, na.rm=TRUE)
> m_eeworkexp1b <- mean(dat1b$eeworkexp, na.rm=TRUE)
> dat1bg <- dat1b[!duplicated(dat1b$erid),]
> m_erexp1b <- mean(dat1bg$erexp, na.rm=TRUE)
>
> sd_score1b <- sd(dat1b$score, na.rm=TRUE)
> sd_eesalary1b <- sd(dat1b$eesalary, na.rm=TRUE)
> sd_eeworkexp1b <- sd(dat1b$eeworkexp, na.rm=TRUE)
> sd_erexp1b <- sd(dat1bg$erexp, na.rm=TRUE)
```

ตัวแปรระดับที่ 2 ต้องสร้างข้อมูลใหม่ก่อน ที่ 1 แถว = 1 กลุ่ม

หาส่วนเบี่ยงเบนมาตรฐาน

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนมากๆ ให้เปลี่ยนเป็นคะแนนมาตรฐานทั้ง 4 ตัวแปร

```
> dat1b$score <- (dat1b$score - m_score1b) / sd_score1b
> dat1b$eesalary <- (dat1b$eesalary - m_eesalary1b) / sd_eesalary1b
> dat1b$eeworkexp <- (dat1b$eeworkexp - m_eeworkexp1b) / sd_eeworkexp1b
> dat1b$erexp <- (dat1b$erexp - m_erexp1b) / sd_erexp1b
```

ทำคะแนนมาตรฐาน

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. ทำ Group mean centering

- ตัวแปรที่คาดว่าจะมีความชันสุ่มกับตัวแปรอื่นในโมเดล หรือมีปฏิสัมพันธ์กับตัวแปรหนึ่งเพื่อทำนายอีกตัวแปรหนึ่งในโมเดล ตัวแปรเหล่านี้ต้องพิจารณาว่าจะทำหรือไม่ทำ Group mean centering
- หากเป็นตัวแปรที่ปกติจะไม่ทำ Group mean centering อยู่แล้ว ก็ไม่ต้องทำ เช่น ตัวแปรเวลา หรือเงื่อนไขในตัวแปรการทดลองที่เกิดจากการสุ่มคนเข้าเงื่อนไข
- หากเป็นตัวแปรที่ต้องทำ Group mean centering เราต้องสร้างตัวแปรใหม่ 2 ตัว คือ ค่าเฉลี่ยกลุ่ม และค่าเบี่ยงเบน จากตัวอย่างนี้จะทำ Group mean centering กับตัวแปรระดับที่ 1 ทั้ง 3 ตัวในโมเดล คือ เงินเดือน ประสบการณ์ทำงาน และคะแนน

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. ทำ Group mean centering

- หากตัวแปรที่มาทำ Group mean centering ไม่มีค่าสูญหายเลย เราอาจคำนวณค่าเฉลี่ยกลุ่ม และค่าเบี่ยงเบนตามปกติ แล้วแนบไปกับข้อมูล
- แต่หากตัวแปรที่มาทำ Group mean centering มีค่าสูญหาย เราอาจจะสร้างตัวแปรเปล่าๆ ไว้ เพื่อให้โมเดลแทนค่าสูญหายมาคำนวณให้ระหว่างแทนค่าสูญหายโดยใส่ NA ไว้ก่อน

```
> dat1b$avescore <- NA
> dat1b$aveeesalary <- NA
> dat1b$aveeeworkexp <- NA
>
> dat1b$diffscore <- NA
> dat1b$diffeesalary <- NA
> dat1b$diffeeworkexp <- NA
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. สร้างตัวแปรปฏิสัมพันธ์

- หากทฤษฎีคาดหวังว่าตัวแปรบางคู่ จะมีปฏิสัมพันธ์กัน ไม่ว่าจะเป็นปฏิสัมพันธ์ในระดับเดียวกัน หรือต่างระดับกัน ให้สร้างตัวแปรปฏิสัมพันธ์แยกออกมาต่างหาก
- หากปฏิสัมพันธ์ดังกล่าว ส่วนประกอบของมันไม่มีค่าสูญหาย ก็คูณส่วนประกอบให้เป็นปฏิสัมพันธ์ แล้วแนบไปกับข้อมูลเลย
- หากส่วนประกอบของปฏิสัมพันธ์ดังกล่าวมีค่าสูญหาย ให้ใส่ค่าเป็น NA ทั้งตัวแปรไปก่อน เพื่อให้โมเดลแทนค่าสูญหายคำนวณปฏิสัมพันธ์หลังแทนค่าแล้ว

```
> dat1b$int1 <- NA    # diffeeworkexp * erexp
> dat1b$int2 <- NA    # aveeeworkexp * erexp
```

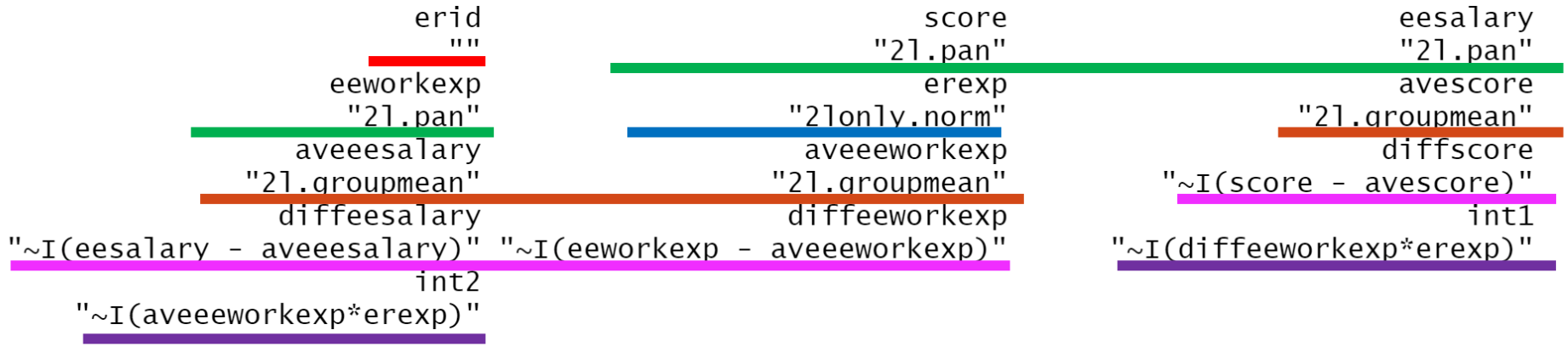
ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. สร้างโมเดลแทนค่าสูญหาย (Imputation Model)

```
> meth1b <- make.method(dat1b)
> meth1b
         erid         score      eesalary     eeworkexp          erexp
           ""         "pmm"         "pmm"         "pmm"         "pmm"
     avescore   aveeesalary  aveeeworkexp     diffscore  diffeesalary
     "logreg"      "logreg"      "logreg"      "logreg"      "logreg"
diffeeworkexp          int1          int2
     "logreg"      "logreg"      "logreg"
```

```
> pred1b <- make.predictorMatrix(dat1b)
> pred1b
```

	erid	score	eesalary	eeworkexp	erexp	avescore	aveeesalary	aveeeworkexp	diffscore	diffeesalary	diffeeworkexp	int1	int2
erid	0	1	1	1	1	1	1	1	1	1	1	1	1
score	1	0	1	1	1	1	1	1	1	1	1	1	1
eesalary	1	1	0	1	1	1	1	1	1	1	1	1	1
eeworkexp	1	1	1	0	1	1	1	1	1	1	1	1	1
erexp	1	1	1	1	0	1	1	1	1	1	1	1	1
avescore	1	1	1	1	1	0	1	1	1	1	1	1	1
aveeesalary	1	1	1	1	1	1	0	1	1	1	1	1	1
aveeeworkexp	1	1	1	1	1	1	1	0	1	1	1	1	1
diffscore	1	1	1	1	1	1	1	1	0	1	1	1	1
diffeesalary	1	1	1	1	1	1	1	1	1	0	1	1	1
diffeeworkexp	1	1	1	1	1	1	1	1	1	1	0	1	1
int1	1	1	1	1	1	1	1	1	1	1	1	0	1
int2	1	1	1	1	1	1	1	1	1	1	1	1	0



1. `erid` เป็นตัวแปร ID ของระดับที่ 2 ไม่มีค่าสูญหาย จึงกำหนดให้เป็น `""`
2. `score, eesalary, eeworkexp` เป็นตัวแปรที่มีค่าสูญหายระดับที่ 1 เป็นตัวแปรต่อเนื่อง จึงใช้ `"2l.pan"`
3. `erexp` เป็นตัวแปรที่มีค่าสูญหาย ระดับที่ 2 เป็นตัวแปรต่อเนื่อง จึงใช้ `"2lonly.norm"`
4. `avescore,aveeesalary,aveeeworkexp` เป็นตัวแปรค่าเฉลี่ยกลุ่ม จึงใช้ `"2l.groupmean"`
5. `diffscore,diffeesalary,diffeeworkexp` เป็นตัวแปรค่าเบี่ยงเบนจากค่าเฉลี่ยกลุ่ม จึงใช้ `~I()` แล้วคำนวณค่าเบี่ยงเบนข้างใน
6. `int1,int2` เป็นตัวแปรปฏิสัมพันธ์ จึงบอกกับโปรแกรมว่า ให้คำนวณโดยตรงจากการคูณตัวแปร จึงใช้ `~I()` แล้วคำนวณปฏิสัมพันธ์ข้างใน

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

```
> meth1b[c("score", "eesalary", "eeworkexp")] <- "2l.pan"
> meth1b["erexp"] <- "2lonly.norm"
> meth1b[c("avescore", "aveeesalary", "aveeeworkexp")] <- "2l.groupmean"
> meth1b["diffscore"] <- "~I(score - avescore)"
> meth1b["diffeesalary"] <- "~I(eesalary - aveeesalary)"
> meth1b["diffeeworkexp"] <- "~I(eeworkexp - aveeeworkexp)"
> meth1b["int1"] <- "~I(diffeeworkexp*erexp)"
> meth1b["int2"] <- "~I(aveeeworkexp*erexp)"
```

Variable	erid	score	eesalary	eeworkexp	erexp	avescore	aveeesalary	aveeeworkexp	diffscore	diffeesalary	diffeeworkexp	int1	int2
erid	0	0	0	0	0	0	0	0	0	0	0	0	0
score	-2	0	0	0	1	0	1	1	0	2	2	1	1
eesalary	-2	0	0	0	1	1	0	1	2	0	2	1	1
eeworkexp	-2	0	0	0	1	1	1	0	2	2	0	0	0

1. ตัวแปร erid เป็นตัวแปร ID ไม่มีค่าสูญหาย ดังนั้นทั้งแถวเป็น 0
2. ตัวแปร score เป็นตัวแปรระดับที่ 1
   1) ถูกทำนายด้วยตัวแปรค่าเบี่ยงเบนระดับที่ 1 คือ diffeesalary, diffeeworkexp ซึ่งเป็นค่าเบี่ยงเบนทำให้ไม่ต้องใส่ค่าเฉลี่ยกลุ่ม แต่คาดหวังความชันสุ่ม จึงใส่เลข 2
   2) ถูกทำนายด้วยตัวแปรระดับที่ 2 คือ erexp และค่าเฉลี่ยของตัวแปรระดับที่ 1 คือ aveeesalary, aveeeworkexp จึงให้ใส่เลข 1
   3) ถูกทำนายด้วยตัวแปรปฏิสัมพันธ์ ซึ่งใส่ 1 เสมอ เพื่อไม่ให้ซ้ำซ้อนกับตัวแปรหลัก
   4) มี erid เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2

Variable	erid	score	eesalary	eeworkexp	erexp	avescore	aveeesalary	aveeeworkexp	diffscore	diffeesalary	diffeeworkexp	int1	int2
erid	0	0	0	0	0	0	0	0	0	0	0	0	0
score	-2	0	0	0	1	0	1	1	0	2	2	1	1
eesalary	-2	0	0	0	1	1	0	1	2	0	2	1	1
eeworkexp	-2	0	0	0	1	1	1	0	2	2	0	0	0

3. ตัวแปร eesalary เป็นตัวแปรระดับที่ 1 คล้าย score จึงใส่ค่าดังนี้ avescore (1), aveeeworkexp (1), diffscore (2), diffeeworkexp (2), erexp (1), int1 (1), int2 (1), erid (-2)
4. ตัวแปร eeworkexp เป็นตัวแปรระดับที่ 1 คล้าย score จึงใส่ค่าดังนี้ avescore (1), aveeesalary (1), diffscore (2), diffeesalary (2), erexp (1), erid (-2) ที่ไม่ใส่ int1 และ int2 เพราะตัวแปรนี้ใช้คำนวณปฏิสัมพันธ์

Variable	erid	score	eesalary	eeworkexp	erexp	avescore	aveeesalary	aveeeworkexp	diffscore	diffeesalary	diffeeworkexp	int1	int2
erexp	-2	0	0	0	0	1	1	1	0	0	0	0	0
avescore	-2	2	0	0	0	0	0	0	0	0	0	0	0
aveeesalary	-2	0	2	0	0	0	0	0	0	0	0	0	0
aveeeworkexp	-2	0	0	2	0	0	0	0	0	0	0	0	0

5. ตัวแปร erexp เป็นตัวแปรระดับที่ 2
   1) ถูกทำนายด้วยตัวแปรระดับที่ 1 ที่แปลงเป็นค่าเฉลี่ยคือ avescore, aveeesalary, aveeeworkexp เพราะเป็นการทำนายตรงๆ ไม่มีอิทธิพลสุ่ม หรือค่าเฉลี่ยกลุ่มมาทำนายทำให้ใส่เลข 1 (จะใส่ score, eesalary, eeworkexp ก็ได้ผลไม่ต่างกัน)
   2) มี erid เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2
6. ตัวแปร avescore, aveeesalary, aveeeworkexp เป็นค่าเฉลี่ยกลุ่มจากตัวแปรระดับที่ 1 ที่ตัวแปรกลุ่มคือ erid ดังนั้น จึงใส่ตัวแปรระดับที่ 1 ของตัวแปรแต่ละตัวเป็น 2 และใส่ erid เป็น -2 นอกจากนั้นเป็น 0

Variable	erid	score	eesalary	eeworkexp	erexp	avescore	aveeesalary	aveeeworkexp	diffscore	diffeesalary	diffeeworkexp	int1	int2
diffscore	0	0	0	0	0	0	0	0	0	0	0	0	0
diffeesalary	0	0	0	0	0	0	0	0	0	0	0	0	0
diffeeworkexp	0	0	0	0	0	0	0	0	0	0	0	0	0
int1	0	0	0	0	0	0	0	0	0	0	0	0	0
int2	0	0	0	0	0	0	0	0	0	0	0	0	0

7. ตัวแปร `diffscore`, `diffeesalary`, `diffeeworkexp`, `int1`, `int2` ใส่คำสั่งคำนวณใน method แล้ว ในเมทริกซ์การทำนาย เลยไม่ต้องใส่ตัวแปรอะไร ใส่ให้เป็น 0 ทั้งหมด

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. สร้างโมเดลแทนค่าสูญหาย (Imputation Model)

```
pred1b[,] <- 0
pred1b[c("score", "eesalary", "eeworkexp", "erexp"), "erid"] <- -2
pred1b[c("avescore", "aveeesalary", "aveeeworkexp"), "erid"] <- -2
pred1b["score", c("aveeesalary", "aveeeworkexp", "erexp")] <- 1
pred1b["score", c("diffeesalary", "diffeeworkexp")] <- 2
pred1b["score", c("int1", "int2")] <- 1
pred1b["eesalary", c("avescore", "aveeeworkexp", "erexp")] <- 1
pred1b["eesalary", c("diffscore", "diffeeworkexp")] <- 2
pred1b["eesalary", c("int1", "int2")] <- 1
pred1b["eeworkexp", c("avescore", "aveeesalary", "erexp")] <- 1
pred1b["eeworkexp", c("diffscore", "diffeesalary")] <- 2
pred1b["erexp", c("avescore", "aveeesalary", "aveeeworkexp")] <- 1
pred1b["avescore", "score"] <- 2
pred1b["aveeesalary", "eesalary"] <- 2
pred1b["aveeeworkexp", "eeworkexp"] <- 2
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

5. กำหนดรอบของการแทนค่าสูญหาย

เนื่องจากตัวแปรค่าเฉลี่ยกลุ่มและค่าเบี่ยงเบน ควรมีการอัพเดทหลักจากมีการแทนค่าตัวแปรหลัก และเนื่องจากตัวแปรปฏิสัมพันธ์ควรมีการอัพเดทค่า ทุกครั้งที่มีการแทนค่าสูญหายให้กับตัวแปรหลักของมัน กล่าวคือ `intexp` ควรมีการเปลี่ยนค่าทุกครั้ง เมื่อค่า `erexp` หรือ `eeworkexp` เปลี่ยนแปลง

```
visit1b <- c("score", "avescore", "diffscore",
             "eesalary", "aveeesalary", "diffeesalary",
             "erexp", "int1", "int2",
             "eeworkexp", "aveeeworkexp", "diffeeworkexp",
             "int1", "int2")
```

ตัวแปรที่อัพเดท หลังตัวแปรที่เกี่ยวข้องมีการแทนค่าสูญหายแล้ว

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

6. ดำเนินการแทนค่าสูญหาย

```
> imp1b <- mice(dat1b, pred=pred1b, meth=meth1b, m=10, maxit=10,
+               visit=visit1b, seed=123321, allow.na = TRUE)
```

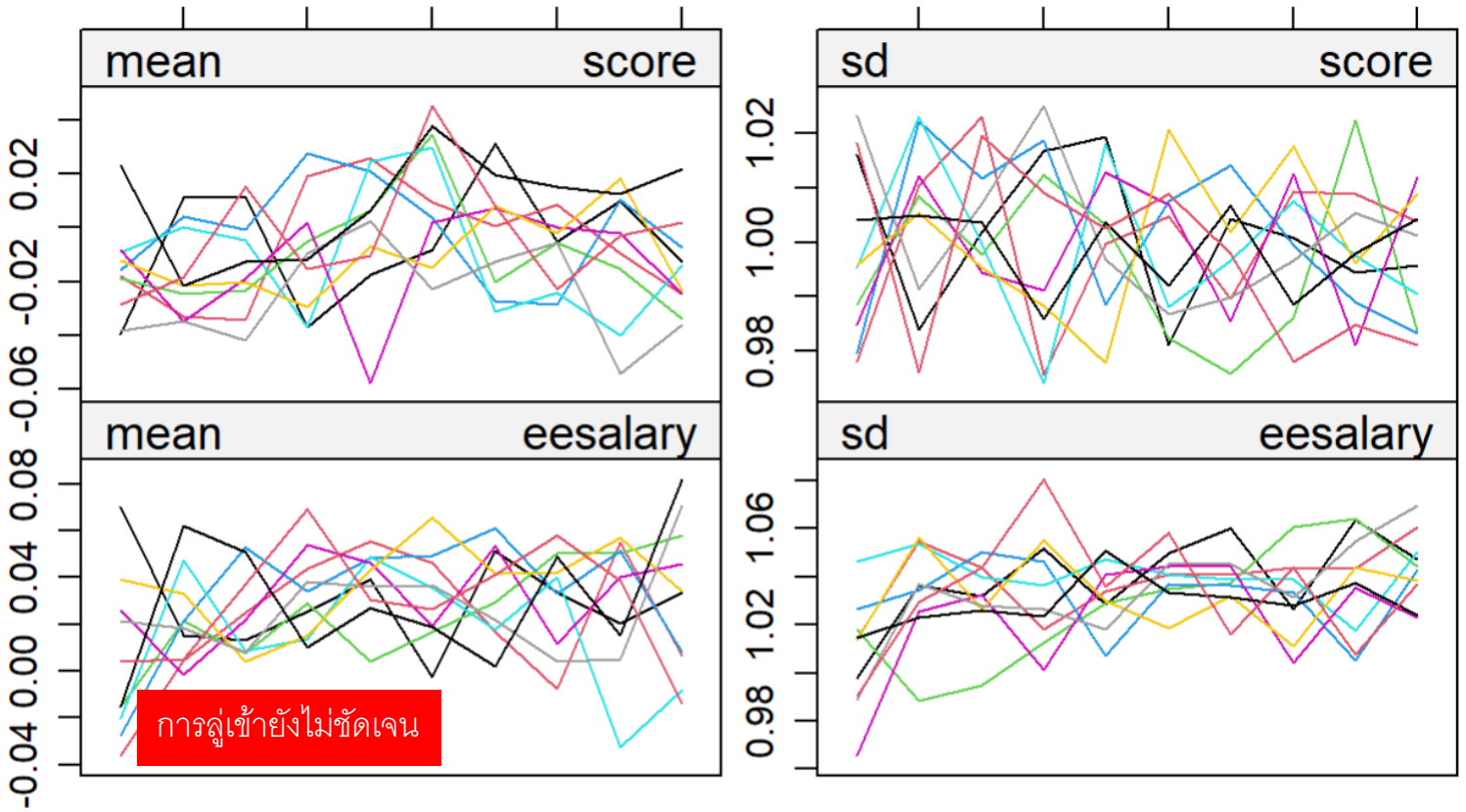
กำหนดลำดับในการแทนค่า

อนุญาตให้ตัวแปรที่มี NA เพียงอย่างเดียวในโมเดล ซึ่งในที่นี้เราทำแล้วในตัวแปรปฏิสัมพันธ์

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. ตรวจสอบการลู่เข้าของการแทนค่า

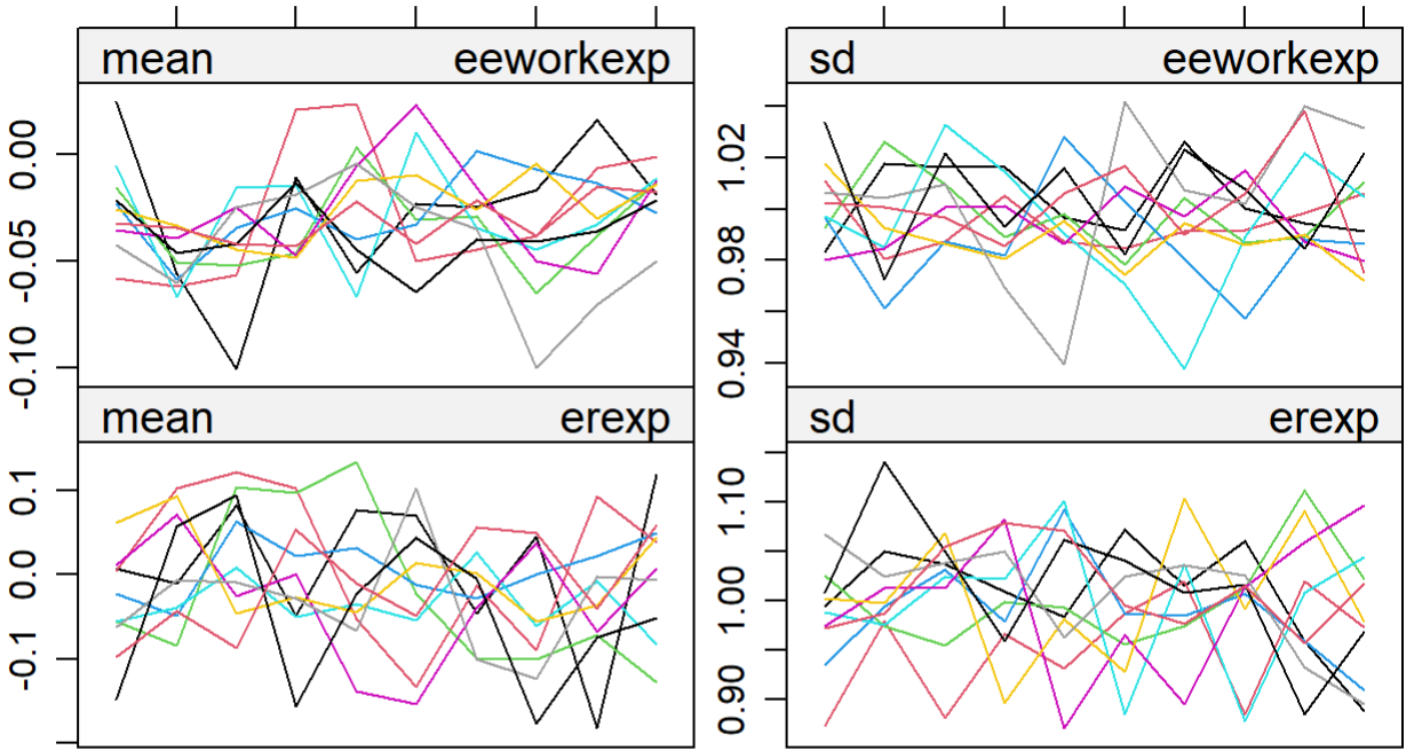
```
> plot(imp1b, c("score", "eesalary"))
```



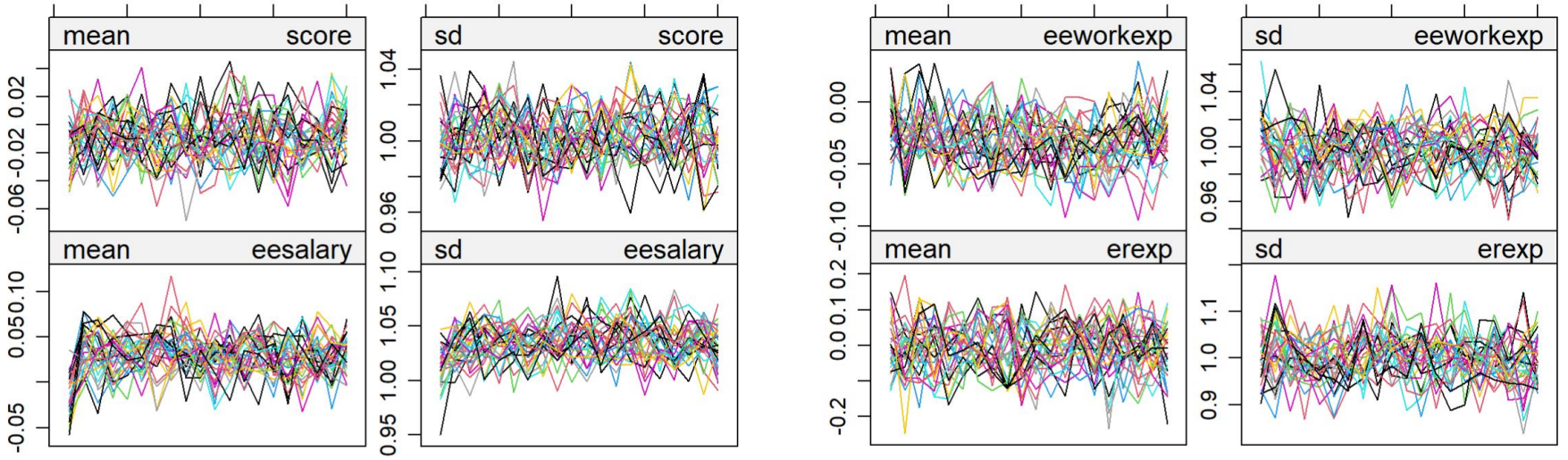
ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. ตรวจสอบการลู่เข้าของการแทนค่า

```
> plot(imp1b, c("eeworkexp", "erexp"))
```



เพิ่มเป็น 30 ข้อมูล วน 20 รอบ



ดูดีขึ้นกว่าเดิม แม้จะมีแยกตัวแตกๆ ออกมาบ้าง

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

8. แปลงจากคะแนนมาตรฐาน ให้กลับเป็นคะแนนปกติ

สร้างข้อมูลทุกชุดที่มีการแทนค่าแล้วออกมา

```
> impdat1b <- complete(imp1b, "long", include = TRUE)
> impdat1b$score <- sd_score1b*impdat1b$score + m_score1b
> impdat1b$eesalary <- sd_eesalary1b*impdat1b$eesalary + m_eesalary1b
> impdat1b$eeworkexp <- sd_eeworkexp1b*impdat1b$eeworkexp + m_eeworkexp1b
> impdat1b$erexp <- sd_erexp1b*impdat1b$erexp + m_erexp1b
> imp1b2 <- as.mids(impdat1b)
```

แปลงกลับเป็นคะแนนดิบ

แปลงกลับให้เป็นผลลัพธ์จาก `mice`

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

9. วิเคราะห์ข้อมูลที่แทนค่าแต่ละชุด

```
> library(lme4)
> fit1b1 <- with(imp1b2, lmer(score ~ eesalary + eeworkexp
+                             + erexp + (1|erid), REML=FALSE))
> fit1b2 <- with(imp1b2, lmer(score ~ eesalary + eeworkexp
+                             + erexp + eeworkexp*erexp+ (1|erid), REML=FALSE))
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

หา Fixed Effect

```
> out1b1 <- pool(fit1b1)
> summary(out1b1)
         term    estimate  std.error   statistic          df      p.value
1 (Intercept) 72.01912482 0.19207913 374.9450829 1207.18482 0.000000e+00
2    eesalary -0.04625989 0.07890441  -0.5862777   94.07149 5.590931e-01
3   eeworkexp  0.26138119 0.01532055  17.0608221  130.80579 0.000000e+00
4       erexp -0.26066300 0.03847539  -6.7747984  660.37816 2.760547e-11
> out1b2 <- pool(fit1b2)
> summary(out1b2)
             term      estimate   std.error   statistic          df      p.value
1     (Intercept) 72.0123968578 0.201746114 356.9456459 835.19551 0.000000e+00
2        eesalary -0.0462388535 0.078948130  -0.5856865  93.90433 5.594912e-01
3       eeworkexp  0.2624386863 0.017298279  15.1713757 157.84359 0.000000e+00
4           erexp -0.2581058459 0.044041512  -5.8605128 402.31522 9.616124e-09
5 eeworkexp:erexp -0.0004085889 0.003173233  -0.1287611 198.81678 8.976769e-01
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

หา Random Effect

```
> source("mimlmtools.R")
> ranefMI(fit1b1)
            (Intercept)
(Intercept)     15.5027
attr(,"stddev")
(Intercept)
   3.937347
attr(,"correlation")
            (Intercept)
(Intercept)           1
> ranefMI(fit1b2)
            (Intercept)
(Intercept)    15.50281
attr(,"stddev")
(Intercept)
    3.93736
attr(,"correlation")
            (Intercept)
(Intercept)           1
> sigmaMI(fit1b1)
[1] 17.84121
> sigmaMI(fit1b2)
[1] 17.84003
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

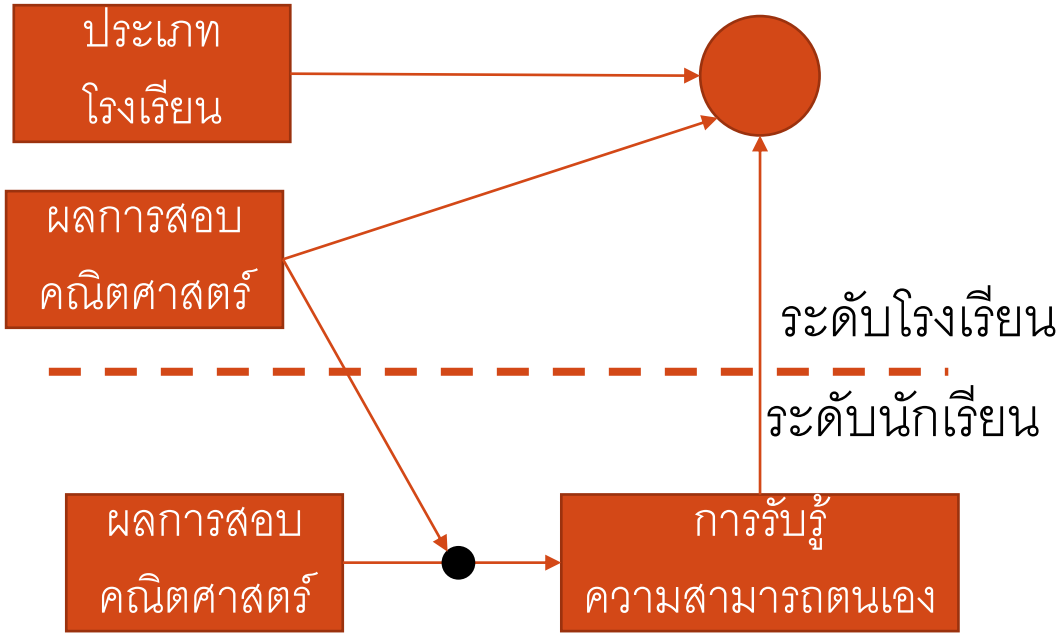
เปรียบเทียบสองโมเดล

```
> anovaMI(fit1b1, fit1b2)
       F      df1      df2      p.F
   0.152    1.000 2575.354    0.697
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

ทำนายการรับรู้ความสามารถของตนเอง ของนักเรียน ด้วยผลการสอบคณิตศาสตร์ และประเภทของโรงเรียน

ตัวอย่างที่ 3 มีปฏิสัมพันธ์ระหว่างค่าเบี่ยงเบนและค่าเฉลี่ยกลุ่ม



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

วิเคราะห์รูปแบบค่าสูญหาย (Missing Data Pattern)

```
> dat2 <-read.table("lecture11ex2.csv", sep=",", header=TRUE, na.strings="999999")
> library(mice)
> library(miceadds)
> md.pattern(dat2)
     id schoolid private efficacy ach
1166  1        1       1        1   1   0
305   1        1       1        1   0   1
291   1        1       1        0   1   1
91    1        1       1        0   0   2
      0        0       0      382 396 778
```

มี 1,166 คนที่มีข้อมูลครบทุกค่า

ที่เหลือมีค่าสูญหายอย่างน้อย 1 ค่า

มีรูปแบบค่าสูญหายทั้งหมด 3 รูปแบบ

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนมากๆ ให้เปลี่ยนเป็นคะแนนมาตรฐานทั้ง 2 ตัวแปร

```
> m_efficacy2 <- mean(dat2$efficacy, na.rm=TRUE)
> m_ach2 <- mean(dat2$ach, na.rm=TRUE)
> sd_efficacy2 <- sd(dat2$efficacy, na.rm=TRUE)
> sd_ach2 <- sd(dat2$ach, na.rm=TRUE)
> dat2$efficacy <- (dat2$efficacy - m_efficacy2) / sd_efficacy2
> dat2$ach <- (dat2$ach - m_ach2) / sd_ach2
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. สร้างตัวแปรค่าเฉลี่ยกลุ่ม ตัวแปรค่าเบี่ยงเบนกลุ่ม และตัวแปรปฏิสัมพันธ์

เนื่องจาก `ach,efficacy` มีค่าสูญหาย ด้วยนั้นค่าเฉลี่ยกลุ่ม ค่าเบี่ยงเบนกลุ่ม และปฏิสัมพันธ์จะมีการอัพเดททุกครั้งที่แทนค่าสูญหายใหม่ จึงใส่ NA ไปก่อน แล้วให้โปรแกรมคำนวณ

```
> dat2$aveach <- NA
> dat2$aveefficacy <- NA
> dat2$diffach <- NA
> dat2$diffefficacy <- NA
> dat2$intach <- NA
> dat2$inteefficacy <- NA
```

```
aveach*diffach
aveefficacy*diffefficacy
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> meth2 <- make.method(dat2)
> meth2
          id     schoolid     efficacy          ach      private       aveach  aveefficacy      diffach diffefficacy       intach
          ""           ""        "pmm"        "pmm"           ""    "logreg"    "logreg"    "logreg"    "logreg"    "logreg"
 intefficacy
    "logreg"
> pred2 <- make.predictorMatrix(dat2)
> pred2
             id schoolid efficacy ach private aveach aveefficacy diffach diffefficacy intach intefficacy
id            0        1        1   1       1      1           1       1            1      1           1
schoolid      1        0        1   1       1      1           1       1            1      1           1
efficacy      1        1        0   1       1      1           1       1            1      1           1
ach           1        1        1   0       1      1           1       1            1      1           1
private       1        1        1   1       0      1           1       1            1      1           1
aveach        1        1        1   1       1      0           1       1            1      1           1
aveefficacy   1        1        1   1       1      1           0       1            1      1           1
diffach       1        1        1   1       1      1           1       0            1      1           1
diffefficacy  1        1        1   1       1      1           1       1            0      1           1
intach        1        1        1   1       1      1           1       1            1      0           1
intefficacy   1        1        1   1       1      1           1       1            1      1           0
```

```
                    id                  schoolid                 efficacy
                    ""                        ""                 "2l.pan"
                   ach                   private                    aveach
               "2l.pan"                        ""              "2l.groupmean"
            aveefficacy                   diffach                diffefficacy
      "2l.groupmean"         "~I(ach - aveach)"   "~I(efficacy - aveefficacy)"
             intach                 intefficacy
"~I(aveach*diffach)" "~I(aveefficacy*diffefficacy)"
```

1. id, schoolid เป็นตัวแปร ID ไม่มีค่าสูญหาย จึงกำหนดให้เป็น ""
2. efficacy, ach เป็นตัวแปรที่มีค่าสูญหาย ระดับที่ 1 เป็นตัวแปรต่อเนื่อง จึงใช้ "2l.pan"
3. private ไม่มีค่าสูญหาย จึงกำหนดให้เป็น ""
4. aveefficacy,aveach เป็นตัวแปรค่าเฉลี่ยกลุ่ม จึงใช้ "2l.groupmean"

```
                id                schoolid                efficacy
                ""                      ""                    "2l.pan"
               ach                 private                     aveach
           "2l.pan"                      ""                "2l.groupmean"
         aveefficacy                diffach                 diffefficacy
     "2l.groupmean"     "~I(ach - aveach)"   "~I(efficacy - aveefficacy)"
             intach                inteffica
"~I(aveach*diffach)" "~I(aveefficacy*diffefficacy)"
```

5. diffach เป็นตัวแปรค่าเบี่ยงเบน ที่เกิดจากนำ ach มาลบด้วย aveach ตรงๆ จึงกำหนดให้เป็น "~I(ach - aveach)"
6. diffefficacy เป็นตัวแปรค่าเบี่ยงเบน ที่เกิดจากนำ efficacy มาลบด้วย aveefficacy ตรงๆ จึงกำหนดให้เป็น "~I(efficacy - aveefficacy)"
7. intach เป็นตัวแปรปฏิสัมพันธ์ที่เกิดจากการคูณตัวแปรภายในข้อมูล จึงกำหนดให้เป็น "~I(aveach*diffach)"
8. inteffica เป็นตัวแปรปฏิสัมพันธ์ที่เกิดจากการคูณตัวแปรภายในข้อมูล จึงกำหนดให้เป็น "~I(aveefficacy*diffefficacy)"

Variable	id	schoolid	efficacy	ach	private	aveach	aveefficacy	diffach	diffefficacy	intach	intefficacy
id	0	0	0	0	0	0	0	0	0	0	0
schoolid	0	0	0	0	0	0	0	0	0	0	0
efficacy	0	-2	0	0	1	1	0	2	0	1	0

1. ตัวแปร `id, schoolid` เป็นตัวแปร ID ไม่มีค่าสูญหาย ดังนั้นทั้งแถวเป็น 0
2. ตัวแปร `efficacy` เป็นตัวแประระดับที่ 1
   1) ตัวแปร `id` ไม่เกี่ยวข้อง ทำให้ใส่เลข 0
   2) มี `schoolid` เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2
   3) ถูกทำนายด้วยค่าเบี่ยงเบนตัวแประระดับที่ 1 คือ `diffach` ทำให้ใส่เลข 2
   4) ถูกทำนายด้วยตัวแประระดับที่ 2 คือ `private, aveach` ทำให้ใส่เลข 1
   5) ตัวแปร `intach` เป็นตัวแปรที่ทำนาย ค่าสูญหายใน `efficacy` ตรงๆ จึงใส่เป็น 1
   6) ตัวแปร `aveefficacy, diffefficacy, intefficacy` ไม่ใช้ทำนาย `efficacy` จึงใส่เป็น 0

Variable	id	schoolid	efficacy	ach	private	aveach	aveefficacy	diffach	diffefficacy	intach	inteefficacy
ach	0	-2	0	0	1	0	1	0	2	0	1
private	0	0	0	0	0	0	0	0	0	0	0

3. ตัวแปร ach เป็นตัวแปรระดับที่ 1
   1) ตัวแปร id ไม่เกี่ยวข้อง ทำให้ใส่เลข 0
   2) มี schoolid เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2
   3) ในที่นี้จะไม่ใช้ตัวแปร efficacy ในการทำนาย เนื่องจากจะใช้ aveefficacy, diffefficacy, และ inteefficacy แทน จึงใส่เลข 0
   4) ถูกทำนายด้วยตัวแปรระดับที่ 2 คือ private, aveefficacy ทำให้ใส่เลข 1
   5) ตัวแปร inteefficacy เป็นตัวแปรที่ทำนาย ค่าสูญหายใน ach ตรงๆ จึงใส่เป็น 1
   6) ตัวแปร diffefficacy เป็นค่าเบี่ยงเบน ที่คาดหวังว่ามีความชันสุ่ม จึงใส่เป็น 2
4. ตัวแปร private ไม่มีค่าสูญหาย จึงใส่ทั้งแถวเป็น 0

Variable	id	schoolid	efficacy	ach	private	aveach	aveefficacy	diffach	diffefficacy	intach	intefficacy
aveach	0	-2	0	2	0	0	0	0	0	0	0
aveefficacy	0	-2	2	0	0	0	0	0	0	0	0
diffach	0	0	0	0	0	0	0	0	0	0	0
diffefficacy	0	0	0	0	0	0	0	0	0	0	0
intefficacy	0	0	0	0	0	0	0	0	0	0	0

5. ตัวแปร `aveefficacy,aveach` เป็นค่าเฉลี่ยกลุ่มจากตัวแปร `efficacy,ach` ที่ตัวแปรกลุ่มคือ `schoolid` ดังนั้น จึงใส่ตัวแปรระดับที่ 1 ที่เกี่ยวข้อง 2 และใส่ `schoolid` เป็น -2 นอกจากนั้นเป็น 0
6. ตัวแปร `diffach, diffefficacy,` และ `intefficacy` ให้กำหนดค่าทุกตัวในแถวเป็น 0 เพราะกำหนดการคำนวณใน method แล้ว

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> meth2[c("efficacy", "ach")] <- "2l.pan"
> meth2[c("aveefficacy", "aveach")] <- "2l.groupmean"
> meth2["diffach"] <- "~I(ach - aveach)"
> meth2["diffefficacy"] <- "~I(efficacy - aveefficacy)"
> meth2["intach"] <- "~I(aveach*diffach)"
> meth2["intefficacy"] <- "~I(aveefficacy*diffefficacy)"
>
> pred2[,] <- 0
> pred2[c("ach", "efficacy", "aveach", "aveefficacy"), "schoolid"] <- -2
> pred2["efficacy", c("aveach", "diffach", "intach", "private")] <- c(1, 2, 1, 1)
> pred2["ach", c("aveefficacy", "diffefficacy", "intefficacy", "private")] <- c(1, 2, 1, 1)
> pred2["aveach", "ach"] <- 2
> pred2["aveefficacy", "efficacy"] <- 2
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. กำหนดรอบของการแทนค่าสูญหาย

ตัวแปรค่าเฉลี่ยกลุ่ม ค่าเบี่ยงเบน และปฏิสัมพันธ์ ควรมีการอัพเดททุกครั้งหลังจาก ที่ `efficacy,ach` มีการแทนค่าแล้ว ดังนั้นจึงกำหนดรอบให้เป็นดังนี้

```
> visit2 <- c("efficacy", "aveefficacy", "diffefficacy", "intefficacy",
+             "ach", "aveach", "diffach", "intach")
```

ตัวแปรที่อัพเดท หลังตัวแปรที่เกี่ยวข้องมีการแทนค่าสูญหายแล้ว

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

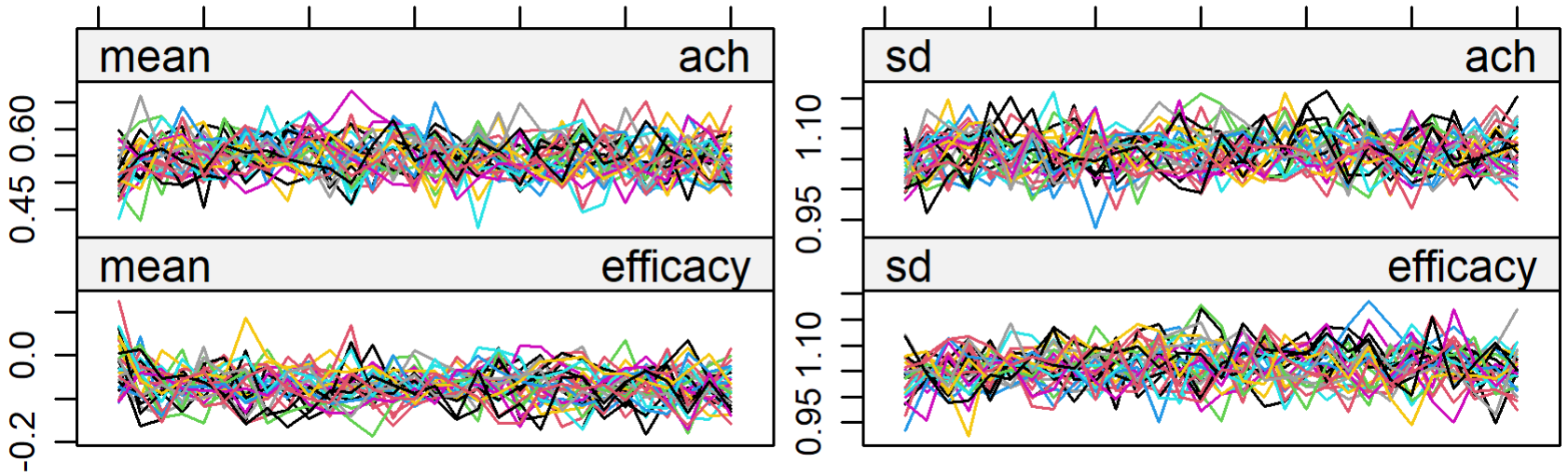
5. ดำเนินการแทนค่าสูญหาย

```
> imp2 <- mice(dat2, pred=pred2, meth=meth2, m=30, maxit=30,
+              visit=visit2, seed=123321, allow.na = TRUE)
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

6. ตรวจสอบการลู่เข้าของการแทนค่า

```
> plot(imp2, c("ach", "efficacy"))
```



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. แปลงจากคะแนนมาตรฐาน ให้กลับเป็นคะแนนปกติ

```
> impdat2 <- complete(imp2, "long", include = TRUE)
> impdat2$efficacy <- sd_efficacy2*impdat2$efficacy + m_efficacy2
> impdat2$ach <- sd_ach2*impdat2$ach + m_ach2
> imp22 <- as.mids(impdat2)
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

8. วิเคราะห์ข้อมูลที่แทนค่าแต่ละข้อมูล

```
> fit21 <- with(imp22, lmer(efficacy ~ diffach + aveach
+                     + private + (1 + diffach|schoolid),
+                     REML=FALSE, control = lmerControl(optimizer ="Nelder_Mead")))
> fit22 <- with(imp22, lmer(efficacy ~ diffach*aveach
+                     + private + (1 + diffach|schoolid),
+                     REML=FALSE, control = lmerControl(optimizer ="Nelder_Mead")))
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

9. รวมผลการวิเคราะห์ข้อมูล

หา Fixed Effect

```
> out21 <- pool(fit21)
> summary(out21)
         term   estimate std.error   statistic       df      p.value
1 (Intercept) 37.0860001 0.3199434 115.9142594 1722.931 0.0000000000
2     diffach  1.2338155 0.3602146   3.4252234 1750.274 0.0006284798
3      aveach  1.1684097 0.3197878   3.6537030 1489.375 0.0002674502
4     private  0.1473936 0.4641962   0.3175243 1698.032 0.7508848763
> out22 <- pool(fit22)
> summary(out22)
            term   estimate std.error   statistic       df      p.value
1    (Intercept) 37.1123680 0.3190892 116.3071782 1725.054 0.000000e+00
2        diffach  1.0012860 0.3066535   3.2652032 1609.513 1.116839e-03
3         aveach  0.9495162 0.3184291   2.9818767 1742.560 2.904615e-03
4        private  0.1483600 0.4642424   0.3195744 1697.642 7.493303e-01
5 diffach:aveach  1.8992530 0.4108074   4.6232201 1622.080 4.077580e-06
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

9. รวมผลการวิเคราะห์ข้อมูล

หา Random Effect

```
> sigmaMI(fit21)
[1] 4.04802
> sigmaMI(fit22)
[1] 4.04799
```

```
> source("mimlmtools.R")
> ranefMI(fit21)
             (Intercept)    diffach
(Intercept)   1.6650105 -0.7308239
diffach      -0.7308239  6.1155271
attr(,"stddev")
(Intercept)     diffach
   1.290353    2.472959
attr(,"correlation")
             (Intercept)    diffach
(Intercept)   1.0000000 -0.2290273
diffach      -0.2290273  1.0000000
> ranefMI(fit22)
             (Intercept)    diffach
(Intercept)   1.6381932 -0.5039789
diffach      -0.5039789  4.1604352
attr(,"stddev")
(Intercept)     diffach
   1.279919    2.039714
attr(,"correlation")
             (Intercept)    diffach
(Intercept)   1.0000000 -0.1930458
diffach      -0.1930458  1.0000000
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

9. รวมผลการวิเคราะห์ข้อมูล

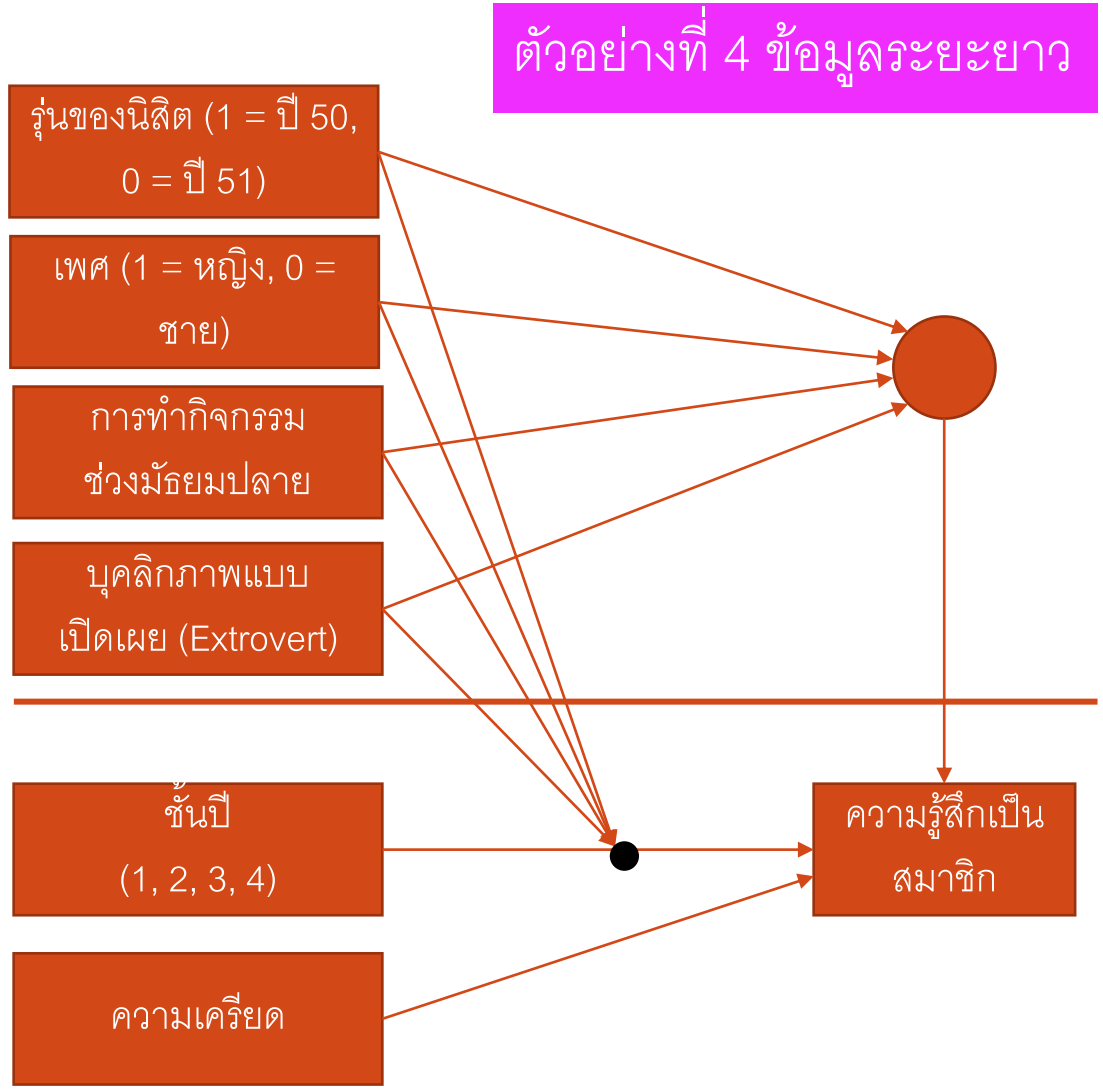
เปรียบเทียบสองโมเดล

```
> anovaMI(fit21, fit22)
          F         df1          df2          p.F
      9.045       1.000 106993.615        0.003
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

ทำนายความรู้สึกเป็นสมาชิก
ของนิสิตจากคณะแห่งหนึ่ง

เก็บข้อมูลจากนิสิต 850 คน เป็นนิสิตรุ่น 2550 ทั้งหมด 426 คน และนิสิตรุ่น 2551 จำนวน 424 คน เก็บข้อมูลระยะยาว ตลอดเวลา 4 ปี



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

วิเคราะห์รูปแบบค่าสูญหาย (Missing Data Pattern)

```
> dat3 <-read.table("lecture11ex3.csv", sep=",", header=TRUE, na.strings="999999")
> md.pattern(dat3)
     rowid pid time cohort female act ext mem stress
2466     1   1    1      1      1   1   1   1      1    0
934      1   1    1      1      1   1   1   0      0    2
         0   0    0      0      0   0   0 934    934 1868
```

มี 2,466 ข้อมูลที่มีข้อมูลครบทุกค่า ที่เหลือมีค่าสูญหายในตัวแปรการเป็นสมาชิกกลุ่มและความเครียดอยู่ 934 ข้อมูล

```
> aggregate(stress ~ time, data = dat3, FUN = function(x) sum(is.na(x)), na.action = NULL)
  time stress
1    1      0
2    2    189
3    3    332
4    4    413
```

ข้อมูลสูญหายมากขึ้นเรื่อยๆ ในเวลาปีที่ 2, 3, และ 4

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนมากๆ ให้เปลี่ยนเป็นคะแนนมาตรฐานทั้ง 2 ตัวแปร

```
> m_mem3 <- mean(dat3$mem, na.rm=TRUE)
> m_stress3 <- mean(dat3$stress, na.rm=TRUE)
> sd_mem3 <- sd(dat3$mem, na.rm=TRUE)
> sd_stress3 <- sd(dat3$stress, na.rm=TRUE)
> dat3$mem <- (dat3$mem - m_mem3) / sd_mem3
> dat3$stress <- (dat3$stress - m_stress3) / sd_stress3
>
> dat3g <- dat3[dat3$time == 1,]
> m_act3 <- mean(dat3g$act, na.rm=TRUE)
> m_ext3 <- mean(dat3g$ext, na.rm=TRUE)
> sd_act3 <- sd(dat3g$act, na.rm=TRUE)
> sd_ext3 <- sd(dat3g$ext, na.rm=TRUE)
> dat3$act <- (dat3$act - m_act3) / sd_act3
> dat3$ext <- (dat3$ext - m_ext3) / sd_ext3
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

2. ย้ายศูนย์กลางตัวแปรเวลา เพื่อให้จุดตัดแกน Y (β_{0j}) มีความหมาย

```
> dat3$time <- dat3$time - 1
```

ผมจะลบ 1 ไปถาวร จนวิเคราะห์ขั้นสุดท้ายเลย แต่ถ้าผู้วิเคราะห์ต้องการย้ายศูนย์กลางใหม่ ก็สามารถทำได้ในภายหลัง

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. สร้างตัวแปรค่าเฉลี่ยกลุ่ม สร้างตัวแปรค่าเบี่ยงเบนกลุ่ม ย้ายศูนย์กลางตัวแปรเวลา และสร้างตัวแปรปฏิสัมพันธ์

สำหรับที่ตัวแปรดั้งเดิม ไม่มีค่าสูญหาย ให้คำนวณตรงๆ เลย

```
> dat3$intact <- dat3$time * dat3$act
> dat3$intext <- dat3$time * dat3$ext
> dat3$intcohort <- dat3$time * dat3$cohort
> dat3$intfemale <- dat3$time * dat3$female
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

3. สร้างตัวแปรค่าเฉลี่ยกลุ่ม สร้างตัวแปรค่าเบี่ยงเบนกลุ่ม ย้ายศูนย์กลางตัวแปรเวลา และสร้างตัวแปรปฏิสัมพันธ์

สำหรับที่ตัวแปรดั้งเดิม มีค่าสูญหาย ให้ใส่ค่า NA ไว้ก่อน

```
> dat3$avemem <- NA
> dat3$diffmem <- NA
> dat3$avestress <- NA
> dat3$diffstress <- NA
> dat3$intmem1 <- NA
> dat3$intmem2 <- NA
> dat3$intstress1 <- NA
> dat3$intstress2 <- NA
```

```
time*avemem
time*diffmem
time*avestress
time*diffstress
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> pred3 <- make.predictorMatrix(dat3)
> pred3
          rowid pid time mem stress cohort female act ext intact intext intcohort intfemale avemem diffmem avestress diffstress intmem1 intmem2 intstress1 intstress2
rowid         0   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
pid           1   0    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
time          1   1    0   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
mem           1   1    1   0      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
stress        1   1    1   1      0      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
cohort        1   1    1   1      1      0      1   1   1      1      1         1         1      1       1         1          1       1       1          1          1
female        1   1    1   1      1      1      0   1   1      1      1         1         1      1       1         1          1       1       1          1          1
act           1   1    1   1      1      1      1   0   1      1      1         1         1      1       1         1          1       1       1          1          1
ext           1   1    1   1      1      1      1   1   0      1      1         1         1      1       1         1          1       1       1          1          1
intact        1   1    1   1      1      1      1   1   1      0      1         1         1      1       1         1          1       1       1          1          1
intext        1   1    1   1      1      1      1   1   1      1      0         1         1      1       1         1          1       1       1          1          1
intcohort     1   1    1   1      1      1      1   1   1      1      1         0         1      1       1         1          1       1       1          1          1
intfemale     1   1    1   1      1      1      1   1   1      1      1         1         0      1       1         1          1       1       1          1          1
avemem        1   1    1   1      1      1      1   1   1      1      1         1         1      0       1         1          1       1       1          1          1
diffmem       1   1    1   1      1      1      1   1   1      1      1         1         1      1       0         1          1       1       1          1          1
avestress     1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         0          1       1       1          1          1
diffstress    1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          0       1       1          1          1
intmem1       1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       0       1          1          1
intmem2       1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       0          1          1
intstress1    1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          0          1
intstress2    1   1    1   1      1      1      1   1   1      1      1         1         1      1       1         1          1       1       1          1          0
> meth3 <- make.method(dat3)
> meth3
     rowid        pid       time        mem     stress     cohort     female        act        ext     intact     intext  intcohort  intfemale     avemem    diffmem
        ""         ""         ""      "pmm"      "pmm"         ""         ""         ""         ""         ""         ""         ""         ""  "logreg"  "logreg"
 avestress diffstress    intmem1    intmem2 intstress1 intstress2
  "logreg"   "logreg"   "logreg"   "logreg"   "logreg"   "logreg"
```

rowid
pid
time

ทั้งแถวเท่ากับ 0

mem
stress

สีแดง คือ แถวที่ต้องมีการปรับค่า

cohort
female
act
ext
intact
intext
intcohort
intfemale

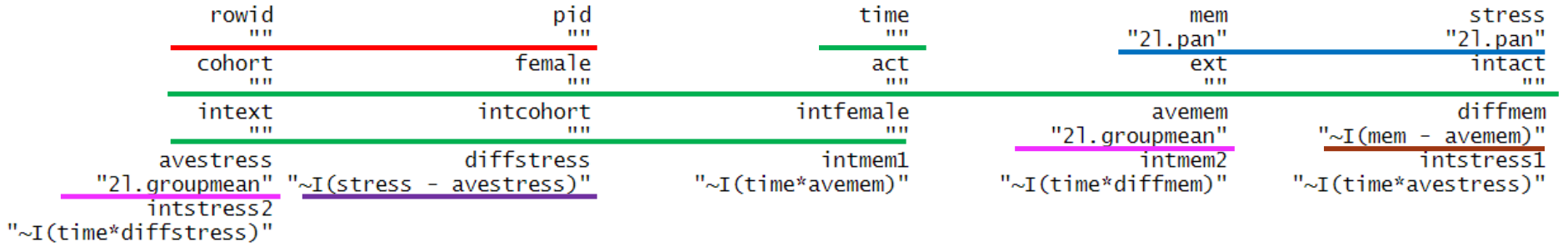
ทั้งแถวเท่ากับ 0

avemem
diffmem
avestress

ทั้งแถวเท่ากับ 0

diffstress
intmem1
intmem2
intstress1
intstress2

ทั้งแถวเท่ากับ 0



1. rowid, pid เป็นตัวแปร ID ไม่มีค่าสูญหาย จึงกำหนดให้เป็น ""
2. time, cohort, female, act, ext, intact, intext, intcohort, intfemale ไม่มีค่าสูญหาย จึงกำหนดให้เป็น ""
3. mem, stress เป็นตัวแปรที่มีค่าสูญหาย ระดับที่ 1 เป็นตัวแปรต่อเนื่อง จึงใช้ "2l.pan"
4. avemem, avestress เป็นตัวแปรค่าเฉลี่ยกลุ่ม จึงใช้ "2l.groupmean"
5. diffmem เป็นตัวแปรค่าเบี่ยงเบน กำหนดให้เป็น "~I(mem - avemem)"
6. diffstress เป็นตัวแปรค่าเบี่ยงเบน กำหนดให้เป็น "~I(stress - avestress)"

```
      rowid         pid        time         mem      stress
         ""          ""          ""    "2l.pan"    "2l.pan"
     cohort      female          act         ext       int
         ""          ""          ""          ""        ""
     intext   intcohort   intfemale     avemem    diffmem
         ""          ""          ""  "2l.groupmean" "~I(mem - avemem)"
  avestress  diffstress    intmem1     intmem2  intstress1
"2l.groupmean" "~I(stress - avestress)" "~I(time*avemem)" "~I(time*diffmem)" "~I(time*avestress)"
  intstress2
"~I(time*diffstress)"
```

8. intmem1 เป็นตัวแปรปฏิสัมพันธ์ กำหนดให้เป็น "~I(time*avemem)"
9. intmem2 เป็นตัวแปรปฏิสัมพันธ์ กำหนดให้เป็น "~I(time*diffmem)"
10. intstress1 เป็นตัวแปรปฏิสัมพันธ์ กำหนดให้เป็น "~I(time*avestress)"
11. intstress2 เป็นตัวแปรปฏิสัมพันธ์ กำหนดให้เป็น "~I(time*diffstress)"

Variable	rowid	pid	time	mem	stress	cohort	female	act	ext	intact	intext	intcohort	intfemale	avemem	diffmem	avestress	diffstress	intmem1	intmem2	intstress1	intstress2
mem	0	-2	2	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1
stress	0	-2	2	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
avemem	0	-2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
avestress	0	-2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1. ตัวแปร mem เป็นตัวแปรระดับที่ 1
   1) ตัวแปร rowid ไม่เกี่ยวข้อง ทำให้ใส่เลข 0
   2) มี pid เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2
   3) ตัวแปร time ใส่เลข 2 เนื่องจากมีความชันสุ่ม
   4) ถูกทำนายด้วยตัวแปรระดับที่ 2 คือ cohort, female, act, ext ทำให้ใส่เลข 1
   5) ตัวแปร stress ไม่ใช้ในการทำนาย เพราะมี avestress และ diffstress แล้วจึงใส่ 0
   6) ตัวแปร avestress และ diffstress ใส่เลข 1
   7) ตัวแปรปฏิสัมพันธ์ทั้งหมด ที่ไม่เกี่ยวกับ mem ให้ทำนาย mem ให้หมด ใส่เลข 1

Variable	rowid	pid	time	mem	stress	cohort	female	act	ext	intact	intext	intcohort	intfemale	avemem	diffmem	avestress	diffstress	intmem1	intmem2	intstress1	intstress2
mem	0	-2	2	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1
stress	0	-2	2	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
avemem	0	-2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
avestress	0	-2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2. ตัวแปร `stress` เป็นตัวแปรระดับที่ 1
   1) ตัวแปร `rowid` ไม่เกี่ยวข้อง ทำให้ใส่เลข 0
   2) มี `pid` เป็นตัวแปร ID กลุ่ม ทำให้ใส่เลข -2
   3) ตัวแปร `time` ใส่เลข 2 เนื่องจากมีความชันสุ่ม
   4) ถูกทำนายด้วยตัวแปรระดับที่ 2 คือ `cohort`, `female`, `act`, `ext` ทำให้ใส่เลข 1
   5) ตัวแปร `mem` ไม่ใช้ในการทำนาย เพราะมี `avemem` และ `diffmem` แล้วจึงใส่ 0
   6) ตัวแปร `avemem` และ `diffmem` ใส่เลข 1
   7) ตัวแปรปฏิสัมพันธ์ทั้งหมด ที่ไม่เกี่ยวกับ `stress` ให้ทำนาย `stress` ให้หมด ใส่เลข 1

Variable	rowid	pid	time	mem	stress	cohort	female	act	ext	intact	intext	intcohort	intfemale	avemem	diffmem	avestress	diffstress	intmem1	intmem2	intstress1	intstress2
mem	0	-2	2	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1
stress	0	-2	2	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
avemem	0	-2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
avestress	0	-2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3. ตัวแปร `avemem` เป็นค่าเฉลี่ยกลุ่มจากตัวแปร `mem` ที่ตัวแปรกลุ่มคือ `pid` ดังนั้น จึงใส่ `mem` เป็น 2 ใส่ `pid` เป็น -2 นอกจากนั้นเป็น 0 ส่วนการจะบอกโปรแกรมว่าให้ค่าเฉลี่ยกลุ่ม จะไปบอกในเวคเตอร์สถิติที่ใช้อีกที
4. ตัวแปร `avestress` เป็นค่าเฉลี่ยกลุ่มจากตัวแปร `stress` ที่ตัวแปรกลุ่มคือ `pid` ดังนั้น จึงใส่ `stress` เป็น 2 ใส่ `pid` เป็น -2 นอกจากนั้นเป็น 0 ส่วนการจะบอกโปรแกรมว่าให้ค่าเฉลี่ยกลุ่ม จะไปบอกในเวคเตอร์สถิติที่ใช้อีกที

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

4. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
pred3[,] <- 0
pred3["mem", c("pid", "time", "cohort", "female",
               "act", "ext", "avestress", "diffstress", "intstress1",
               "intstress2", "intact", "intext", "intcohort", "intfemale")] <- c(-2, 2, 1, 1,
                                                                                 1, 1, 1, 1, 1,
                                                                                 1, 1, 1, 1, 1)
pred3["stress", c("pid", "time", "cohort", "female",
               "act", "ext", "avemem", "diffmem", "intmem1",
               "intmem2", "intact", "intext", "intcohort", "intfemale")] <- c(-2, 2, 1, 1,
                                                                              1, 1, 1, 1, 1,
                                                                              1, 1, 1, 1, 1)

pred3["avemem", c("pid", "mem")] <- c(-2, 2)
pred3["avestress", c("pid", "stress")] <- c(-2, 2)
meth3[c("mem", "stress")] <- "2l.pan"
meth3[c("avemem", "avestress")] <- "2l.groupmean"
meth3["diffmem"] <- "~I(mem - avemem)"
meth3["diffstress"] <- "~I(stress - avestress)"
meth3["intmem1"] <- "~I(time*avemem)"
meth3["intmem2"] <- "~I(time*diffmem)"
meth3["intstress1"] <- "~I(time*avestress)"
meth3["intstress2"] <- "~I(time*diffstress)"
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

5. กำหนดรอบของการแทนค่าสูญหาย

ตัวแปรค่าเฉลี่ยกลุ่ม ค่าเบี่ยงเบน และปฏิสัมพันธ์ ควรมีการอัพเดททุกครั้งหลังจากที่ efficacy มีการแทนค่าแล้ว ดังนั้นจึงกำหนดรอบให้เป็นดังนี้

mem → avemem → diffmem → intmem1 → intmem2 →
stress → avestress → diffstress → intstress1 → intstress2

```
> visit3 <- c("mem", "avemem", "diffmem", "intmem1", "intmem2",
+             "stress", "avestress", "diffstress", "intstress1", "intstress2")
```

ตัวแปรที่อัพเดท หลังตัวแปรที่เกี่ยวข้องมีการแทนค่าสูญหายแล้ว

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

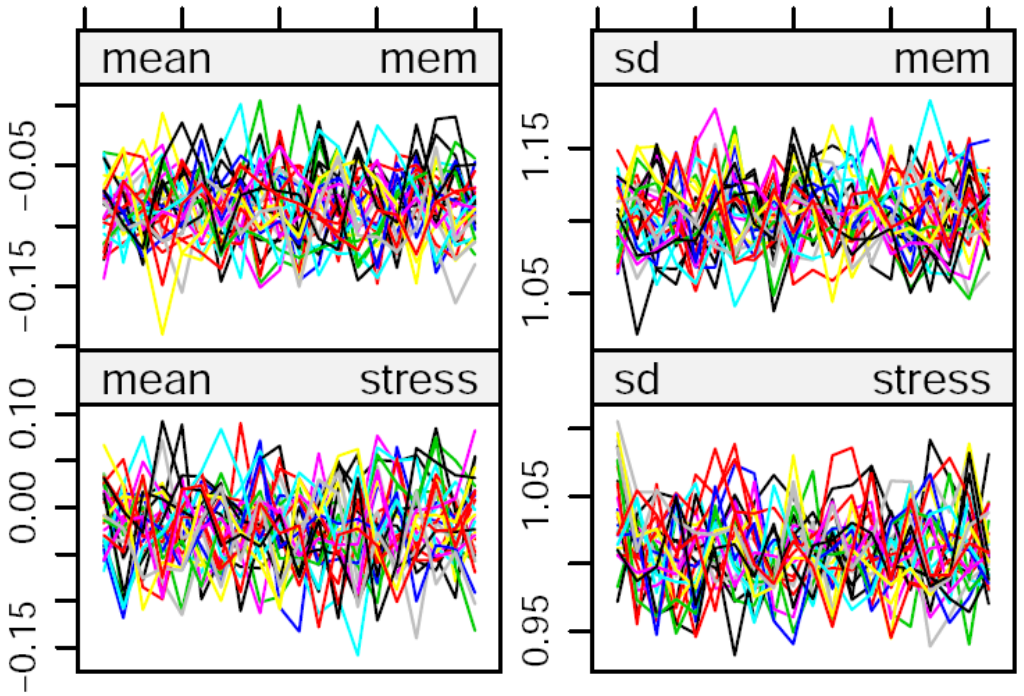
6. ดำเนินการแทนค่าสูญหาย

```
> imp3 <- mice(dat3, pred=pred3, meth=meth3, m=30, maxit=20,
+              visit=visit3, seed=123321, allow.na = TRUE)
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

7. ตรวจสอบการลู่เข้าของการแทนค่า

```
> plot(imp3, c("mem", "stress"))
```



ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

8. แปลงจากคะแนนมาตรฐาน ให้กลับเป็นคะแนนปกติ

```
> impdat3 <- complete(imp3, "long", include = TRUE)
> impdat3$mem <- sd_mem3*impdat3$mem + m_mem3
> impdat3$stress <- sd_stress3*impdat3$stress + m_stress3
> impdat3$act <- sd_act3*impdat3$act + m_act3
> impdat3$ext <- sd_ext3*impdat3$ext + m_ext3
> imp32 <- as.mids(impdat3)
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

9. วิเคราะห์ข้อมูลที่แทนค่าแต่ละข้อมูล

```
> fit31 <- with(imp32, lmer(mem ~ time + (1|pid), REML=FALSE,
+                           control = lmerControl(optimizer ="Nelder_Mead")))
> fit32 <- with(imp32, lmer(mem ~ time + (1 + time|pid), REML=FALSE,
+                           control = lmerControl(optimizer ="Nelder_Mead")))
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

หา Fixed Effect

```
> out31 <- pool(fit31)
> summary(out31)
               estimate std.error statistic        df p.value
(Intercept) 61.989459 0.6234996  99.42181 3175.6393       0
time        -1.880244 0.1799847 -10.44669  295.3776       0
>
> out32 <- pool(fit32)
> summary(out32)
               estimate std.error  statistic        df      p.value
(Intercept) 61.989459 0.5431517 114.129173 3053.4635 0.000000e+00
time        -1.880244 0.2284058  -8.232032  664.4133 8.881784e-16
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

หา Random Effect

```
> source("mimlmtools.R")
> ranefMI(fit31)
            (Intercept)
(Intercept)    255.6447
attr(,"stddev")
(Intercept)
   15.98889
attr(,"correlation")
            (Intercept)
(Intercept)           1
> ranefMI(fit32)
            (Intercept)      time
(Intercept)    205.3823  1.34630
time             1.3463 25.21274
attr(,"stddev")
(Intercept)        time
  14.331166    5.021229
attr(,"correlation")
            (Intercept)       time
(Intercept)  1.00000000 0.01870898
time         0.01870898 1.00000000
> sigmaMI(fit31)
[1] 97.28151
> sigmaMI(fit32)
[1] 55.26033
```

ตัวอย่างการวิเคราะห์ด้วยการแทนค่าแบบพหุ

10. รวมผลการวิเคราะห์ข้อมูล

เปรียบเทียบสองโมเดล

```
> anovaMI(fit31, fit32)
        F      df1      df2      p.F
116.309    2.000 787.755    0.000
```

การเปลี่ยนโครงสร้างข้อมูล

- ในการวัดข้อมูลระยะยาว หากแต่ละคนมีเวลาที่ถูกวัดเหมือนกัน เช่น 1, 2, 3 เดือนเหมือนกัน เพียงแต่บางคนอาจมีค่าสูญหายบางเดือน การจัดโครงสร้างข้อมูลจะทำได้สองแบบ คือ แบบกว้าง (Wide format) และแบบยาว (Long format)
- ขณะที่ หากเวลาที่ถูกวัดแต่ละคนแตกต่างกัน (Individually varying time) เช่น คนที่ 1 วัดเมื่อ 2, 7, 14 เดือน คนที่ 2 วัดเมื่อ 3, 4, 6, 9, 12 เดือน ฯลฯ รูปแบบข้อมูลแบบที่ใช้ในการวิเคราะห์พหุระดับที่ผ่านมาหรือที่เรียกว่าแบบยาว (Long format) เป็นเพียงรูปแบบเดียวที่ทำได้

ข้อมูลแบบยาว (Long Format)

L2ID	T	Y	X	W	Q
1	1	2	9	2	9
1	2	3	8	2	9
1	3	4	7	2	9
2	1	3	8	3	8
2	2	4	7	3	8
2	3	5	6	3	8
3	1	4	7	4	7
3	2	5	6	4	7
3	3	6	5	4	7
4	1	5	6	5	6
4	2	6	5	5	6
4	3	7	4	5	6

ข้อมูลแบบกว้าง (Wide Format)

L2ID	Y1	Y2	Y3	X1	X2	X3	W	Q
1	2	3	4	9	8	7	2	9
2	3	4	5	8	7	6	3	8
3	4	5	6	7	6	5	4	7
4	5	6	7	6	5	4	5	6

T = ตัวแปรเวลา

Y, X = ตัวแปรที่เปลี่ยนแปลงตามเวลา

W, Q = ตัวแปรที่ไม่เปลี่ยนแปลงตามเวลา

ข้อมูลแบบยาว (Long Format)

L2ID	T	Y	X	W	Q
1	1	2	9	2	-
1	2	3	8	2	-
1	3	4	7	2	-
2	1	3	8	3	8
2	2	-	-	3	8
2	3	-	-	3	8
3	1	4	7	4	7
3	2	5	6	4	7
3	3	6	5	4	7
4	1	5	6	-	6
4	2	6	5	-	6
4	3	-	-	-	6

ข้อมูลแบบกว้าง (Wide Format)

L2ID	Y1	Y2	Y3	X1	X2	X3	W	Q
1	2	3	4	9	8	7	2	-
2	3	-	-	8	-	-	3	8
3	4	5	6	7	6	5	4	7
4	5	6	-	6	5	-	-	6

ในกรณีการจัดการค่าสูญหาย สามารถเปลี่ยนโครงสร้างข้อมูล 2 ระดับ (แบบยาว) มาเป็นข้อมูลระดับเดียว (แบบกว้าง) แล้วแทนค่าสูญหายได้

เมื่อแทนค่าสูญหายแล้ว ก็เปลี่ยนกลับให้กลายเป็นข้อมูลแบบยาวเหมือนเดิมเพื่อวิเคราะห์พหุระดับ

แทนค่าด้วยข้อมูลแบบกว้างดีกว่า ตรงที่ความสัมพันธ์ตัวเอง (Autocorrelation) จะถูกนำมาคิดในการแทนค่าสูญหายโดยอัตโนมัติ

การเปลี่ยนโครงสร้างข้อมูล

```
> dat3 <-read.table("lecture10ex3.csv", sep=",", header=TRUE, na.strings="999999")
> dat3 <- dat3[,-1]
> psych::describe(dat3)
```

นำตัวแปร `rowid` ออก เพื่อทำให้ตัวอย่างดูสะอาดขึ้น

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
pid	1	3400	425.50	245.41	425.5	425.50	315.05	1	850	849	0.00	-1.20	4.21
time	2	3400	2.50	1.12	2.5	2.50	1.48	1	4	3	0.00	-1.36	0.02
mem	3	2466	59.61	18.34	59.0	59.61	19.27	3	113	110	0.00	-0.29	0.37
stress	4	2466	50.45	13.57	50.0	50.39	13.34	3	99	96	0.05	0.18	0.27
cohort	5	3400	0.50	0.50	1.0	0.50	0.00	0	1	1	0.00	-2.00	0.01
female	6	3400	0.50	0.50	0.0	0.49	0.00	0	1	1	0.02	-2.00	0.01
act	7	3400	50.22	13.12	50.5	50.29	14.08	9	87	78	-0.06	-0.46	0.23
ext	8	3400	49.53	9.73	49.0	49.47	8.90	23	82	59	0.10	0.00	0.17

เปลี่ยนข้อมูลจากแบบยาวให้เป็นแบบกว้าง ด้วยคำสั่ง `reshape`

ตัวแปร ID ระดับที่ 2

ตัวแปรที่มีค่าแตกต่างกันระหว่างเวลา

```
> dat3w <- reshape(data=dat3, idvar="pid", timevar="time", v.names=c("mem", "stress"), direction="wide")
```

ตัวแปรเวลา

เปลี่ยนเป็นข้อมูลแบบกว้าง

```
> head(dat3, 12)
   pid time mem stress cohort female act ext
1    1    1  68     66      0      0  34  50
2    1    2  45     53      0      0  34  50
3    1    3  43     72      0      0  34  50
4    1    4  41     56      0      0  34  50
5    2    1  61     35      0      0  31  55
6    2    2  65     56      0      0  31  55
7    2    3  51     51      0      0  31  55
8    2    4  40     42      0      0  31  55
9    3    1  77     55      0      0  31  56
10   3    2  60     47      0      0  31  56
11   3    3  NA     NA      0      0  31  56
12   3    4  NA     NA      0      0  31  56
```

ตัวเลขของเวลาถูกใส่ท้ายชื่อของตัวแปร

```
> head(dat3w)
   pid cohort female act ext mem.1 stress.1 mem.2 stress.2 mem.3 stress.3 mem.4 stress.4
1    1      0      0  34  50    68       66    45       53    43       72    41       56
5    2      0      0  31  55    61       35    65       56    51       51    40       42
9    3      0      0  31  56    77       55    60       47    NA       NA    NA       NA
13   4      0      0  29  36    70       59    51       61    39       49    32       60
17   5      0      0  56  65    80       62    67       60    52       52    63       76
21   6      0      0  41  47    78       66    66       58    55       44    49       52
```

การเปลี่ยนโครงสร้างข้อมูล

```
> library(mice)
> md.pattern(dat3w)
    pid cohort female act ext mem.1 stress.1 mem.2 stress.2 mem.3 stress.3 mem.4 stress.4
437   1      1      1   1   1     1        1     1        1     1        1     1        1    0
81    1      1      1   1   1     1        1     1        1     1        1     0        0    2
143   1      1      1   1   1     1        1     1        1     0        0     0        0    4
189   1      1      1   1   1     1        1     0        0     0        0     0        0    6
      0      0      0   0   0     0        0   189      189   332      332   413      413 1868
```

437 คน ตอบข้อมูลทั้ง 4 เวลา

189 คน ออกจากการเก็บข้อมูลในการวัดครั้งที่ 2

143 คน ออกจากการเก็บข้อมูลในการวัดครั้งที่ 3

81 คน ออกจากการเก็บข้อมูลในการวัดครั้งที่ 4

การเปลี่ยนโครงสร้างข้อมูล

1. เพื่อการประมาณค่าด้วยโปรแกรมได้รวดเร็ว ไม่มีปัญหาจากที่คอมพิวเตอร์คูณเลขจำนวนมากๆ ให้เปลี่ยนเป็นคะแนนมาตรฐานทั้ง 2 ตัวแปร
   ข้อมูลระดับเดียว ไม่ได้ใช้การคำนวณจำนวนมาก จึงไม่แปลงคะแนนมาตรฐาน
2. ย้ายศูนย์กลางตัวแปรเวลา
   ไม่มีตัวแปรข้อมูลเวลา ในข้อมูลแบบกว้าง
3. สร้างตัวแปรค่าเฉลี่ยกลุ่ม สร้างตัวแปรค่าเบี่ยงเบนกลุ่ม ย้ายศูนย์กลางตัวแปรเวลา และสร้างตัวแปรปฏิสัมพันธ์
   ตัวแปรค่าเฉลี่ยกลุ่ม และค่าเบี่ยงเบนกลุ่มไม่มีในข้อมูลระดับกว้างอยู่แล้ว ส่วนตัวแปรปฏิสัมพันธ์ระหว่างเวลาและตัวแปรอื่น จะถูกจัดการโดยอัตโนมัติ เพราะช่วงเวลาต่างๆ ได้ถูกใช้แยกตัวแปรในข้อมูลแบบกว้างแล้ว

การเปลี่ยนโครงสร้างข้อมูล

4. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> pred3w <- make.predictorMatrix(dat3w)
> pred3w
         pid cohort female act ext mem.1 stress.1 mem.2 stress.2 mem.3 stress.3 mem.4 stress.4
pid        0      1      1   1   1     1        1     1        1     1        1     1        1
cohort     1      0      1   1   1     1        1     1        1     1        1     1        1
female     1      1      0   1   1     1        1     1        1     1        1     1        1
act        1      1      1   0   1     1        1     1        1     1        1     1        1
ext        1      1      1   1   0     1        1     1        1     1        1     1        1
mem.1      1      1      1   1   1     0        1     1        1     1        1     1        1
stress.1   1      1      1   1   1     1        0     1        1     1        1     1        1
mem.2      1      1      1   1   1     1        1     0        1     1        1     1        1
stress.2   1      1      1   1   1     1        1     1        0     1        1     1        1
mem.3      1      1      1   1   1     1        1     1        1     0        1     1        1
stress.3   1      1      1   1   1     1        1     1        1     1        0     1        1
mem.4      1      1      1   1   1     1        1     1        1     1        1     0        1
stress.4   1      1      1   1   1     1        1     1        1     1        1     1        0
> meth3w <- make.method(dat3w)
> meth3w
     pid   cohort   female      act      ext    mem.1 stress.1    mem.2 stress.2    mem.3 stress.3    mem.4 stress.4
      ""       ""       ""       ""       ""       ""        ""    "pmm"    "pmm"    "pmm"    "pmm"    "pmm"    "pmm"
```

การเปลี่ยนโครงสร้างข้อมูล

pid
cohort
female
act
ext
mem.1
stress.1

ทั้งแถวเท่ากับ 0 เพราะไม่มีค่าสูญหาย

mem.2
stress.2
mem.3
stress.3
mem.4
stress.4

ให้แทนค่าสูญหาย

ทำนายด้วยตัวแปรอื่น นอกจากตัวแปร pid และตัวเอง

	pid	cohort	female	act	ext	mem.1	stress.1	mem.2	stress.2	mem.3	stress.3	mem.4	stress.4
mem.2	0	1	1	1	1	1	1	0	1	1	1	1	1
mem.3	0	1	1	1	1	1	1	1	1	0	1	1	1
mem.4	0	1	1	1	1	1	1	1	1	1	1	0	1
stress.2	0	1	1	1	1	1	1	1	0	1	1	1	1
stress.3	0	1	1	1	1	1	1	1	1	1	0	1	1
stress.4	0	1	1	1	1	1	1	1	1	1	1	1	0

การเปลี่ยนโครงสร้างข้อมูล

4. สร้างโมเดลในการแทนค่าสูญหาย (Imputation Model)

```
> pred3w[,] <- 0
> pred3w[c("mem.2", "mem.3", "mem.4", "stress.2", "stress.3", "stress.4"), ] <- 1
> pred3w[, "pid"] <- 0
> diag(pred3w) <- 0
>
> meth3w[c("mem.2", "mem.3", "mem.4", "stress.2", "stress.3", "stress.4")] <- "pmm"
```

โมเดลที่ใช้แทนค่าสูญหายในแต่ละตัวแปร ใช้ predictive mean matching (pmm) ซึ่งเป็นการทำนายค่าสูญหายในตัวแปรต่างๆ ไม่ว่าจะเป็นตัวแปรต่อเนื่อง หรือดัมมี่ โดยไม่สร้างข้อตกลงเบื้องต้นว่าตัวแปรตามมีการกระจายรูปแบบใด

การเปลี่ยนโครงสร้างข้อมูล

5. กำหนดรอบของการแทนค่าสูญหาย

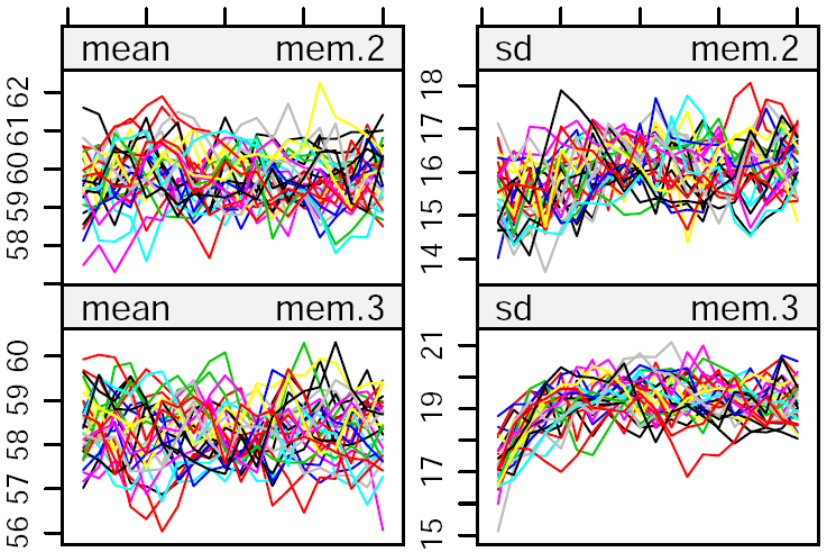
   ไม่ต้องทำ เพราะไม่มีตัวแปรที่สร้างขึ้นมานอกเหนือจากข้อมูลที่มี

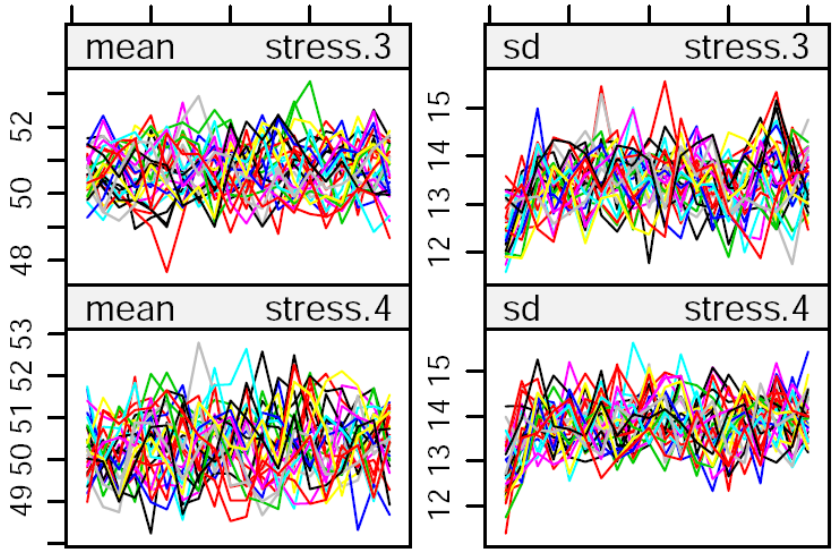
6. ดำเนินการแทนค่าสูญหาย

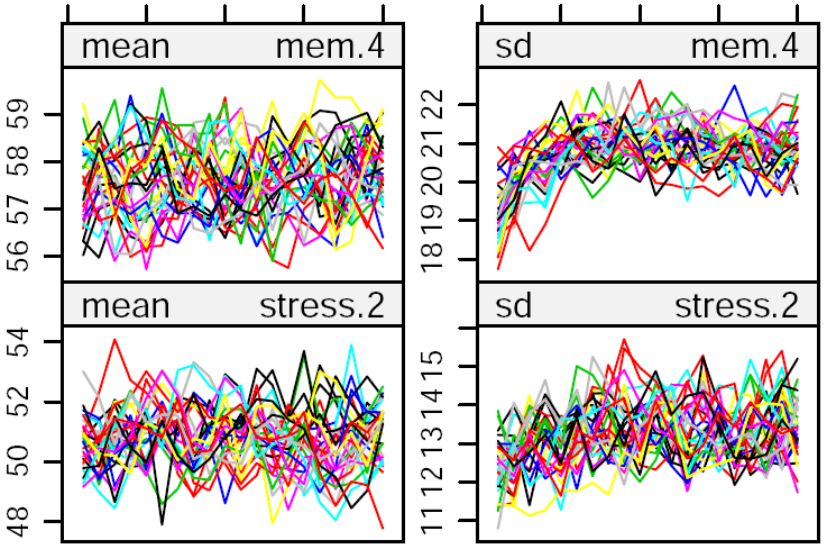
```
> imp3w <- mice(dat3w, pred=pred3w, meth=meth3w, m=30, maxit=20, seed=123321)
```

7. ตรวจสอบการลู่เข้าของการแทนค่า

```
> plot(imp3w, c("mem.2", "mem.3"))
> plot(imp3w, c("mem.4", "stress.2"))
> plot(imp3w, c("stress.3", "stress.4"))
```







ดูข้อมูลลู่เข้าค่อนข้างดี

การเปลี่ยนโครงสร้างข้อมูล

8. แปลงข้อมูลแบบกว้าง ให้เป็นข้อมูลแบบยาวเหมือนเดิม

สร้างข้อมูลที่แทนค่าแล้ว ซ้อนกันเป็นข้อมูลชุดเดียวกัน โดยใส่ข้อมูลเดิมลงไปด้วย

```
> impdat3w <- complete(imp3w, "long", include = TRUE)
```

นิยามตัวแปร ที่เป็นตัวแปรเดียวกันแต่วัดในเวลาที่ต่างกัน

ลำดับไล่จากเวลาน้อยไปหามาก

```
> mem <- c("mem.1", "mem.2", "mem.3", "mem.4")
> stress <- c("stress.1", "stress.2", "stress.3", "stress.4")
> timevarying <- list(mem, stress)
```

นำข้อมูลแต่ละตัวแปร มาผูกอยู่ใน `list` เดียวกัน

การเปลี่ยนโครงสร้างข้อมูล

8. แปลงข้อมูลแบบกว้าง ให้เป็นข้อมูลแบบยาวเหมือนเดิม

สร้างตัวแปรใหม่ชื่อว่า id2 ซึ่งตัวแปร ID ที่ไม่สนว่าจะเป็นข้อมูลที่แทนค่าแล้วชุดอะไร

```
> impdat3w$id2 <- 1:nrow(impdat3w)
```

เปลี่ยนจากข้อมูลแบบกว้างให้เป็นข้อมูลแบบยาว

ตัวแปร ID ใหม่ที่แต่ละแถวมีค่าต่างกัน

ค่าของตัวแปรเวลา

ชื่อของตัวแปรเวลา

```
> dat3l <- reshape(data=impdat3w, idvar="id2", times=0:3, timevar="time",
+                  varying = timevarying, v.names=c("mem", "stress"), direction="long")
```

ลิสของตัวแปรที่เปลี่ยนแปลงตามเวลา

ตั้งชื่อตัวแปรที่กำหนดไว้ในลิส

เปลี่ยนเป็นข้อมูลแบบยาว

การเปลี่ยนโครงสร้างข้อมูล

8. แปลงข้อมูลแบบกว้าง ให้เป็นข้อมูลแบบยาวเหมือนเดิม

สร้างตัวแปร ID ใหม่ ที่เป็นตัวแปร ID แยกชุดของข้อมูล

```
> dat3l$.id <- ave(dat3l$id2, dat3l$.imp, FUN=seq_along)
```

ใช้คำสั่ง `ave` แต่เปลี่ยนจากการใช้ค่าเริ่มต้น `FUN=mean` เป็น `FUN=seq_along` คำสั่ง `seq_along` เป็นการสร้างลำดับตัวแปร 1, 2, 3, ... ไปเรื่อยๆ ตามความยาวของข้อมูล

เปลี่ยนข้อมูลกลับ เป็นผลลัพธ์ของ `mice`

```
> imp3w2 <- as.mids(dat3l)
```

การเปลี่ยนโครงสร้างข้อมูล

9. วิเคราะห์ข้อมูลที่แทนค่าแต่ละข้อมูล

```
> library(lme4)
> fit3w1 <- with(imp3w2, lmer(mem ~ time + (1|pid), REML=FALSE,
+                             control = lmerControl(optimizer ="Nelder_Mead")))
> fit3w2 <- with(imp3w2, lmer(mem ~ time + (1 + time|pid), REML=FALSE,
+                             control = lmerControl(optimizer ="Nelder_Mead")))
+
```

การเปลี่ยนโครงสร้างข้อมูล

10. รวมผลการวิเคราะห์ข้อมูล

หา Fixed Effect

```
> out3w1 <- pool(fit3w1)
> summary(out3w1)
              estimate std.error statistic        df     p.value
(Intercept) 61.525388 0.6291355 97.793538 804.36339 0.000000000
time        -1.299043 0.3258106 -3.987112  75.49626 0.000153171
> out3w2 <- pool(fit3w2)
> summary(out3w2)
              estimate std.error  statistic       df      p.value
(Intercept) 61.525388 0.5992798 102.665552 693.5058 0.0000000000
time        -1.299043 0.3555780  -3.653329 106.2591 0.0004038423
```

การเปลี่ยนโครงสร้างข้อมูล

10. รวมผลการวิเคราะห์ข้อมูล

หา Random Effect

```
> source("mimlmtools.R")
> ranefMI(fit3w1)
            (Intercept)
(Intercept)    156.6986
attr(,"stddev")
(Intercept)
   12.51793
attr(,"correlation")
            (Intercept)
(Intercept)           1
> ranefMI(fit3w2)
            (Intercept)       time
(Intercept)    155.6958 -15.46970
time           -15.4697  25.86102
attr(,"stddev")
(Intercept)        time
  12.477812    5.085373
attr(,"correlation")
            (Intercept)       time
(Intercept)   1.0000000 -0.2437927
time         -0.2437927  1.0000000
> sigmaMI(fit3w1)
[1] 179.4665
> sigmaMI(fit3w2)
[1] 136.3648
```

การเปลี่ยนโครงสร้างข้อมูล

10. รวมผลการวิเคราะห์ข้อมูล

เปรียบเทียบสองโมเดล

```
> anovaMI(fit3w1, fit3w2)
      F     df1     df2     p.F
11.653   2.000  48.859   0.000
```

คาบต่อไป

- การวิเคราะห์ข้อมูลคู่ (Dyadic Data Analysis)
- การบ้านที่ 10