ข้อตกลงเบื้องต้นก่อนการใช้สถิติ (Statistical Assumptions)

สันทัด พรประเสริฐมานิต

โครงร่างการนำเสนอ (รายการข้อตกลงเบื้องต้นที่เกี่ยวข้อง)

1. สำหรับการประมาณค่าแบบ ML: ค่าคงเหลือทั้งระดับองค์ประกอบและตัวบ่งชี้ มีการกระจายเป็น MVN
2. สำหรับการประมาณค่าทุกรูปแบบ
   A. หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน (Independently Distributed)
   B. ความสัมพันธ์เป็นความสัมพันธ์เชิงเส้น (Linearity)
   C. ตัวแปรอิสระภายนอก (Exogenous Independent Variables) และตัวแปรสังเกตได้ที่ถูกทำให้เป็นตัวแปรตาม (Observed Dependent Variables) ปราศจากความผิดพลาดในการวัด
   D. โมเดลองค์ประกอบเป็นรูปแบบสะท้อน (Reflexive Measurement Models)
   E. ปราศจากค่าสุดโต่ง (Outliers) หรือค่าที่มีอิทธิพลสูง (Influential Cases)
   F. ความแปรปรวนของค่าคงเหลือเท่ากัน (Identically Distributed)
   G. จัดการค่าสูญหายได้ถูกต้อง
   H. ความสัมพันธ์ระหว่างตัวแปรอิสระสูงผิดปกติ (Multicollinearity)
3. สำหรับการแปลความหมายโมเดล : ใส่ตัวแปรที่เกี่ยวข้องในโมเดลทั้งหมด

ข้อตกลงเบื้องต้นก่อนการใช้สถิติ

- ข้อตกลงเบื้องต้นก่อนการใช้สถิติเป็นเงื่อนไขที่นักสถิติใช้ในการสร้างสถิติ ทำเป็นสูตรสำหรับวิเคราะห์ข้อมูล หรือเป็นลักษณะของโมเดล ที่กำหนดว่าความสัมพันธ์ระหว่างตัวแปรทั้งหมดต้องเป็นไปตามที่กำหนด
- หากละเมิดข้อตกลงเบื้องต้น อาจเกิดความผิดพลาดในผลการวิเคราะห์ หรือการตีความหมายโมเดล
- ผลกระทบสามารถแบ่งออกง่ายๆ เป็น 3 กลุ่มด้วยกัน คือ
  - บางข้อ แก้ไขได้ ทั้งใช้สถิติที่พัฒนาขึ้นมาเพื่อแก้ไข หรือการใช้โมเดลที่ถูกต้องสามารถแก้ไขได้
  - บางข้อ ตรวจสอบได้ แต่ถ้าละเมิดแล้วไม่สามารถแก้ไขได้
  - บางข้อ ตรวจสอบไม่ได้ และแก้ไขไม่ได้ แต่ต้องใช้การออกแบบงานวิจัย การเก็บข้อมูล เพื่อช่วยเหลือให้ไม่มีปัญหาในเรื่องข้อตกลงเบื้องต้นนี้

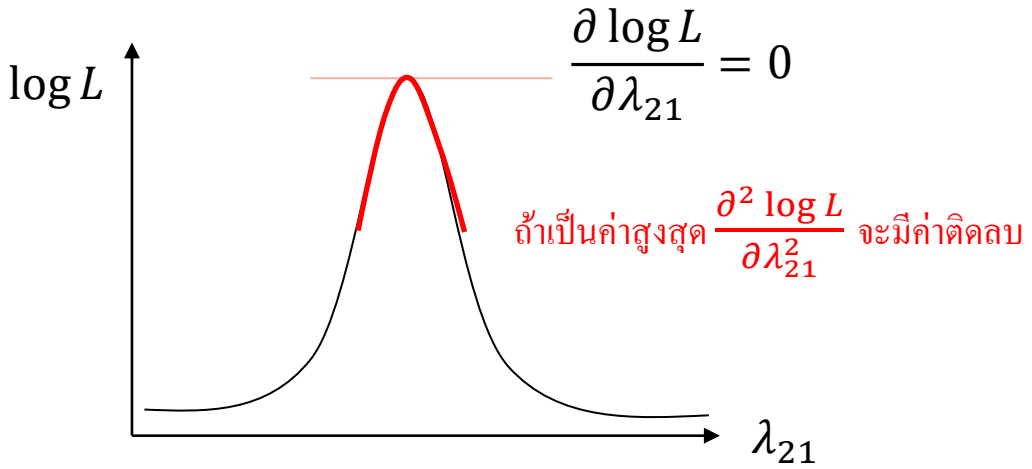
1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- ใน SEM ทั้งแบบตัวแปรต่อเนื่อง e หรือ ζ จะมีเงื่อนไขว่ามีการกระจายเป็น MVN โดยส่วนใหญ่จะส่งผลให้การกระจายของตัวบ่งชี้เป็น MVN ด้วย
- เมื่อการกระจายของตัวบ่งชี้ (หรือการกระจายแบบมีเงื่อนไขกรณีที่มีตัวแปรต้นภายนอก) เป็น MVN แล้ว ค่าของพารามิเตอร์จะถูกประมาณค่าให้มีความเป็นไปได้สูงสุด เมื่อ MVN เป็นจริง

$$\log L = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{N}{2}tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \frac{N}{2}\left[\boldsymbol{m}_y - \boldsymbol{\mu}\right]'\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{m}_y - \boldsymbol{\mu}\right]$$

- ค่าพารามิเตอร์หนึ่งจะมีความเป็นไปได้สูงสุด ก็คือความชัน ของ $\log L$ เมื่อเทียบกับพารามิเตอร์ดังกล่าว จะต้องมีค่าเป็น 0
  - $\frac{\partial \log L}{\partial \theta}$ มีค่าเป็น 0 หมายความว่า θ จุดดังกล่าวจะมีค่า $\log L$ สูงที่สุด (หรือต่ำที่สุด)

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ



- ความโค้งของการเปลี่ยนแปลงของ $\log L$ เมื่อเทียบกับพารามิเตอร์ จะเป็นฟังก์ชั่นของ *SE* ของพารามิเตอร์นั้น ยิ่งโค้งเยอะ ยิ่ง *SE* ต่ำ

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- กล่าวในเชิงเทคนิค ให้ $\boldsymbol{\theta}$ เป็นเวคเตอร์ของพารามิเตอร์ทั้งหมด ที่ $\boldsymbol{\mu}(\boldsymbol{\theta})$ และ $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ ถูกนำไปใช้ในสมการ $\log L$ ข้างต้น หากการกระจายของตัวบ่งชี้ (หรือแบบมีเงื่อนไขตามตัวแปรภายนอก) เป็น MVN แล้ว

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sim \mathrm{MVN}(0, \mathbf{I}^{-1})$$

- เรียก **I** ว่า Information Matrix มีนิยามดังนี้ (Pawitan, 2001)

$$\mathbf{I} = E(-\mathbf{H})|_{\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}} \text{ โดย } \mathbf{H} = \{h_{ij}\} = \left\{\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\right\}$$

- **H** จะเรียกว่า Hessian Matrix ซึ่งเป็น Second-order derivative ของ Log Likelihood นั่นเอง

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

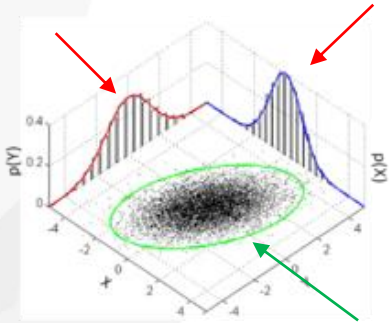
- รากที่สองของค่าสมาชิกแนวทแยงของ $\frac{1}{n}\mathbf{I}^{-1}$ ก็คือ *SE* ของแต่ละพารามิเตอร์
  - $\frac{1}{n}\mathbf{I}^{-1}$ อาจเรียกว่าเมทริกซ์ความแปรปรวนร่วมของค่าสถิติ (Asymptotic Covariance Matrix)
  - จากตัวอย่าง ยิ่ง $\frac{\partial^2 \log L}{\partial \lambda_{21}^2}$ มีค่าติดลบเยอะๆ (โค้งลงแรงๆ) แสดงว่าค่านี้ใน Hessian Matrix ก็จะติดลบเยอะ ค่าของ Information Matrix ก็จะบวกเยอะ ค่าของ Inverse Information Matrix ก็จะมีค่าน้อย ซึ่งหมายความว่า *SE* น้อย
- ความถูกต้องของการทดสอบทางสถิติก็ขึ้นอยู่กับว่า *SE* คำนวณถูกต้องหรือไม่ และการกระจายของพารามิเตอร์ใน Sampling Distribution เป็นโค้งปกติเมื่อ $N \to \infty$ หรือไม่
- หากการกระจายของตัวบ่งชี้ไม่เป็น MVN แล้ว *SE* ของแต่ละพารามิเตอร์จะคำนวณออกมาไม่ถูกต้อง

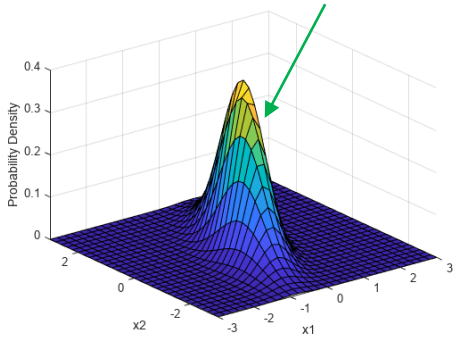
1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

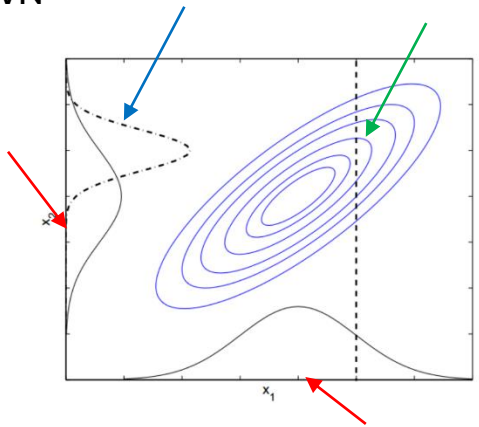
- โดยสรุป เมื่อการประมาณค่าพารามิเตอร์ การหาค่า *SE* รวมถึงการคำนวณ χ^2 มาจาก MVN ทั้งหมด ทำให้เมื่อการกระจายไม่ได้เป็น MVN จะทำให้ผลการคำนวณค่าต่างๆ ผิดพลาด (Finch, West, & MacKinnon, 1997)
  - การประมาณค่าพารามิเตอร์ยังค่อนข้างถูกต้อง ยังไม่ค่อยมีผิดเพี้ยน
  - ค่า *SE* ต่ำกว่าปกติ ก่อให้เกิด Type I error ที่สูงขึ้น
  - ค่า χ^2 ไม่ได้มีการกระจายเป็น Chi-square distribution และมีค่าสูงกว่าปกติ ส่งผลให้ความเหมาะสมของโมเดลแย่กว่าปกติ

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- การตรวจสอบ MVN สามารถตรวจสอบได้หลายมุมมอง
  - มุมมองของ MVN
    - การกระจายตัวแปรแต่ละตัวโดยไม่สนใจตัวแปรอื่น หรือ Marginal Distribution ต้องเป็นโค้งปกติ
    - การกระจายตัวแปรแต่ละตัวเมื่อควบคุมอีกตัวแปรให้คงที่ หรือ Conditional Distribution ต้องเป็นโค้งปกติ
    - การกระจายทุกตัวแปรต้องร่วมกันเป็นโค้งปกติ หรือ Joint Distribution
    - เป็นไปได้ที่ Marginal Distribution เป็นโค้งปกติ แต่ไม่ได้เป็น MVN







1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- การตรวจสอบ MVN สามารถตรวจสอบได้หลายมุมมอง
  - การทดสอบนัยสำคัญ
    - การกระจายของตัวแปรแต่ละตัวต้องเป็นโค้งปกติ ซึ่งสามารถทดสอบว่าแตกต่างจาก 0 หรือไม่ จากนำค่าความเบ้ (Skewness) และความโด่ง (Kurtosis) ไปหารด้วย *SE* แล้วดูว่าน้อยกว่า -1.96 หรือมากกว่า 1.96 ($p < .05$) หรือไม่
    - การทดสอบของ Mardia ว่าความเบ้ในภาพรวมทั้งหมด และความโด่งในภาพรวมทั้งหมดแตกต่างจากโค้งปกติอย่างมีนัยสำคัญหรือไม่
    - อย่างไรก็ตาม ถ้ากลุ่มตัวอย่างสูงมาก การเบี่ยงเบนแค่เล็กน้อยก็ถึงระดับนัยสำคัญ
  - ขนาดของความเบ้และโด่ง West, Finch, & Curran (1995) พบว่าขนาดความเบ้มากกว่า 2 และขนาดความโด่งมากกว่า 7 ถือว่าการกระจายห่างจากโค้งปกติรุนแรง
  - ความโด่งมักส่งผลหนักต่อผลการวิเคราะห์ข้อมูล

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- วิธีการแก้ไข เมื่อตัวแปรไม่ได้เป็นโค้งปกติมีดังนี้
  - ใช้การประมาณค่ารูปแบบอื่นทั้ง WLS (หรือ ADF; Browne, 1984), DWLS, ULS ไม่ได้เหมาว่าการกระจายตัวบ่งชี้เป็น MVN
  - ใช้การปรับค่า Chi-square และ *SE* ให้ถูกต้อง (Satorra & Bentler, 1988, 1994) วิธีการนี้เรียกว่า Scaled Chi-square และ Robust SE โดยการปรับจะคำนึงถึง Multivariate Kurtosis เป็นหลัก วิธีการของ Satorra & Bentler ใน lavaan กำหนดได้โดย estimator="MLM"
    - ในผลการวิเคราะห์จะพบค่า Scaling correction factor ซึ่งคือค่า Chi-square ปกติหารด้วยค่า Scaled Chi-square เช่น 1.75 หมายถึง Chi-square มีค่ามากกว่า Scaled chi-square ถึง 75%
    - Yuan & Bentler (2000) ได้เสนอการปรับในกรณีที่ข้อมูลมีค่าสูญหาย โดยเหมาว่าค่าสูญหายเป็น MCAR (แต่พิสูจน์ว่าคงทนกับ MAR) เรียกว่า estimator="MLR"
    - วิธีการนี้ ต้องการกลุ่มตัวอย่างสูง เช่น > 250 หน่วย (Yu & Muthen, 2002)

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- วิธีการแก้ไข เมื่อตัวแปรไม่ได้เป็นโค้งปกติมีดังนี้
  - ใช้ bootstrap ในการหา *SE* ที่เหมาะสม ซึ่ง bootstrap มีประโยชน์ที่ไม่ได้เหมาว่า Sampling Distribution ของพารามิเตอร์เป็นแบบสมมาตร (เช่น ความแปรปรวนขององค์ประกอบ) แต่ก็มีปัญหาในการหาดัชนีความเหมาะสม ที่ไม่ได้สูตรในการหา χ^2 ที่ชัดเจน ที่นำไปหาดัชนีความเหมาะสมได้
  - การเปรียบเทียบโมเดลที่ซ้อนกันด้วย Chi-square difference test ให้ใช้แนวทางของ Satorra & Bentler (2001)

$$\Delta\chi^2_{SB} = \frac{(\chi^2_0 scf_0 - \chi^2_1 scf_1)(df_0 - df_1)}{df_0 scf_0 - df_1 scf_1}$$

  - โดย χ^2_0 และ χ^2_1 เป็นค่า Chi-square แบบปกติ และ *scf* คือ Scaling correction factor ของแต่ละโมเดล

1: ค่าคงเหลือมีการกระจายเป็นโค้งปกติ

- นักวิจัยหลายคนแนะนำว่าให้ใช้ “MLM” ในการวิเคราะห์ข้อมูลไปเลย ไม่ว่าการกระจายจะเป็น MVN หรือไม่
  - จะทำให้กำลังในการทดสอบทางสถิติตกลงเล็กน้อย
  - เมื่อเทียบกับความเสี่ยงในการผิดข้อตกลงเบื้องต้น ใช้ MLM ไปเลยน่าจะเหมาะสมกว่า
- สรุป
  - กรณีไม่มีค่าสูญหาย ใช้ “MLM”
  - กรณีมีค่าสูญหาย ใช้ “MLR” หรือ “ML” พร้อมกับ missing=“ML”
  - กรณีเป็นตัวแปรแบบจัดกลุ่ม ใช้ “WLSMV”
  - กรณีเป็นตัวแปรแบบจัดกลุ่ม มีค่าสูญหาย พิจารณาจัดการค่าสูญหายด้วย Multiple Imputation

จากตัวอย่างแบบทดสอบเชาวน์ปัญญาย่อย 9 ชุด ของ Holzinger & Swineford (1939)
เป็นข้อมูลดิบที่มีอยู่แล้วใน lavaan

ดู Marginal Distribution

```
> library(lavaan)
> library(psych)
> describe(HolzingerSwineford1939)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
id	1	301	176.55	105.94	163.00	176.78	140.85	1.00	351.00	350.00	-0.01	-1.36	6.11
sex	2	301	1.51	0.50	2.00	1.52	0.00	1.00	2.00	1.00	-0.06	-2.00	0.03
ageyr	3	301	13.00	1.05	13.00	12.89	1.48	11.00	16.00	5.00	0.69	0.20	0.06
agemo	4	301	5.38	3.45	5.00	5.32	4.45	0.00	11.00	11.00	0.09	-1.22	0.20
school*	5	301	1.52	0.50	2.00	1.52	0.00	1.00	2.00	1.00	-0.07	-2.00	0.03
grade	6	300	7.48	0.50	7.00	7.47	0.00	7.00	8.00	1.00	0.09	-2.00	0.03
x1	7	301	4.94	1.17	5.00	4.96	1.24	0.67	8.50	7.83	-0.25	0.31	0.07
x2	8	301	6.09	1.18	6.00	6.02	1.11	2.25	9.25	7.00	0.47	0.33	0.07
x3	9	301	2.25	1.13	2.12	2.20	1.30	0.25	4.50	4.25	0.38	-0.91	0.07
x4	10	301	3.06	1.16	3.00	3.02	0.99	0.00	6.33	6.33	0.27	0.08	0.07
x5	11	301	4.34	1.29	4.50	4.40	1.48	1.00	7.00	6.00	-0.35	-0.55	0.07
x6	12	301	2.19	1.10	2.00	2.09	1.06	0.14	6.14	6.00	0.86	0.82	0.06
x7	13	301	4.19	1.09	4.09	4.16	1.10	1.30	7.43	6.13	0.25	-0.31	0.06
x8	14	301	5.53	1.01	5.50	5.49	0.96	3.05	10.00	6.95	0.53	1.17	0.06
x9	15	301	5.37	1.01	5.42	5.37	0.99	2.78	9.25	6.47	0.20	0.29	0.06

ไม่มีความเบ้ที่ < -2 หรือ > 2

ถ้าเอาความเบ้ ไปหารด้วย se จะพบว่ามี z มีขนาดมากกว่า 1.96 ($p < .05$) จำนวนมาก แต่โดยรวมขนาดความเบ้ไม่สูงจนอันตราย

ไม่มีความโด่งที่ < -7 หรือ > 7

ถ้าเอาความเบ้ ไปหารด้วย se จะพบว่ามี z มีขนาดมากกว่า 1.96 ($p < .05$) จำนวนมาก แต่โดยรวมขนาดความโด่งไม่สูงจนอันตราย

ดู Joint Distribution

```
> library(semTools)
> mardiaSkew(HolzingerSwineford1939[,paste0("x", 1:9)])
         b1d          chi           df            p
6.806892e+00 3.414791e+02 1.650000e+02 2.198164e-14
> mardiaKurtosis(HolzingerSwineford1939[,paste0("x", 1:9)])
         b2d            z            p
102.90374304   2.40658874   0.01610229
```

ดูการทดสอบของ Mardia พบว่าความเบ้และความโด่งต่างจาก MVN อย่างมีนัยสำคัญ

```
> mhs <- '
+ visual  =~ x1 + x2 + x3
+ textual =~ x4 + x5 + x6
+ speed   =~ x7 + x8 + x9
+ '
> outhS <- cfa(mhs, data = HolzingerSwineford1939, estimator="mlm")
> summary(outhS, fit.measures = TRUE)
lavaan 0.6-12 ended normally after 35 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        21

  Number of observations                           301

Model Test User Model:

                                              Standard      Robust
  Test Statistic                                85.306      80.872
  Degrees of freedom                            24          24
  P-value (Chi-square)                          0.000       0.000
  Scaling correction factor                                 1.055
    Satorra-Bentler correction
```

ใช้วิธีของ Satorra and Bentler ในการปรับค่า Chi-square และ *SE*

คอลัมน์ Robust คือ ค่าที่ปรับแล้ว

85.306 / 80.872 = 1.055

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.931       0.925
  Tucker-Lewis Index (TLI)                       0.896       0.887

  Robust Comparative Fit Index (CFI)                         0.932
  Robust Tucker-Lewis Index (TLI)                            0.897

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3737.745   -3737.745
  Loglikelihood unrestricted model (H1)      -3695.092   -3695.092

  Akaike (AIC)                                7517.490    7517.490
  Bayesian (BIC)                              7595.339    7595.339
  Sample-size adjusted Bayesian (BIC)         7528.739    7528.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.092       0.089
  90 Percent confidence interval - lower         0.071       0.068
  90 Percent confidence interval - upper         0.114       0.110
  P-value RMSEA <= 0.05                          0.001       0.001

  Robust RMSEA                                               0.091
  90 Percent confidence interval - lower                     0.070
  90 Percent confidence interval - upper                     0.113

Standardized Root Mean Square Residual:

  SRMR                                           0.065       0.065
```

ค่าดัชนีความเหมาะสมทั้งหมดให้ดูคอลัมน์ด้านขวา

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  visual =~
    x1                1.000
    x2                0.554    0.103    5.359    0.000
    x3                0.729    0.115    6.367    0.000
  textual =~
    x4                1.000
    x5                1.113    0.066   16.762    0.000
    x6                0.926    0.060   15.497    0.000
  speed =~
    x7                1.000
    x8                1.180    0.152    7.758    0.000
    x9                1.082    0.132    8.169    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  visual ~~
    textual           0.408    0.082    4.966    0.000
    speed             0.262    0.055    4.762    0.000
  textual ~~
    speed             0.173    0.055    3.139    0.002
```

ค่า *SE* ในที่นี้เป็นค่าที่ปรับด้วยวิธี Satorra and Bentler เรียบร้อยแล้ว

```
> mhs2 <- '
+ visual  =~ 1*x1 + 1*x2 + 1*x3
+ textual =~ 1*x4 + 1*x5 + 1*x6
+ speed   =~ 1*x7 + 1*x8 + 1*x9
+ '
> ouths2 <- cfa(mhs2, data = HolzingerSwineford1939, estimator="mlm")
> anova(ouths, ouths2)
Scaled Chi-Squared Difference Test (method = “satorra.bentler.2001”)

lavaan NOTE:
    The “Chisq” column contains standard test statistics, not the
    robust test that should be reported per model. A robust difference
    test is a function of two standard (not robust) statistics.

       Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
ouths  24 7517.5 7595.3  85.305
ouths2 30 7527.6 7583.2 107.411     22.334       6   0.001053 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ใช้คำสั่ง anova เพื่อทำ Scaled Chi-square difference test

พบว่าสองโมเดลแตกต่างกันอย่างมีนัยสำคัญ เลือกโมเดลที่น้ำหนักองค์ประกอบแตกต่างกันภายในองค์ประกอบ

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- การที่ข้อมูลเป็นอิสระจากกัน (Independent) คือ ข้อมูลของหน่วยหนึ่ง ไม่สามารถบอกใบ้ข้อมูลของอีกคนหน่วยหนึ่งได้
- การละเมิดข้อตกลงเบื้องต้นนี้ ที่มักเกิดขึ้น คือ
  - ข้อมูลเกิดตามการเวลา เช่น นำข้อมูลสังเกตพฤติกรรมการเริ่มพูดของทารกจาก Case หนึ่ง มาวิเคราะห์ว่าระหว่างช่วงที่ให้แรงเสริมและไม่ให้แรงเสริมมีความถี่ในการพูดแตกต่างกันหรือไม่ การนำมาหา t-test วิเคราะห์แตกต่างอาจไม่เหมาะสม เพราะคะแนนจากเวลาที่ใกล้กันจะคล้ายคลึงกัน มากกว่าเวลาที่ไกลกัน
    - แก้ไขได้โดยการกำหนดโครงสร้างของค่าคงเหลือระหว่างหน่วย ซึ่งเกินขอบเขตของวิชานี้
  - ข้อมูลจากการสุ่มระดับกลุ่ม (Cluster) เช่น เก็บข้อมูลจากหลายโรงเรียน เพื่อตรวจสอบการตั้งเป้าหมายการเรียนแบบเน้นเรียนรู้ (Mastery Orientation) หรือเน้นเป้าหมาย (Goal Orientation) ส่งผลต่อการเรียนอย่างไร นักเรียนที่อยู่โรงเรียนเดียวกันจะคล้ายคลึงกันมากกว่านักเรียนที่อยู่ต่างโรงเรียนกัน

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- กรณีที่ไม่สนใจข้อมูลเวลาหรือกลุ่ม จะส่งผลให้เกิดอคติในการประมาณค่าพารามิเตอร์, *SE*, และค่าความเหมาะสมของโมเดล
- ในบทนี้จะเน้นที่หน่วยตัวอย่างสามารถจับเป็นกลุ่มได้ เช่น นักเรียนซ้อนอยู่ในโรงเรียน (Students nested in Schools) ข้อมูลแบบนี้จะเรียกว่าเป็นข้อมูลพหุระดับ (Multlievel data structure)
  - หากวิเคราะห์ข้อมูลระดับนักเรียน โดยไม่สนใจตัวแปรโรงเรียนเลย จะเรียกว่าการวิเคราะห์แบบไม่สนกลุ่ม (Disaggregated Analysis) และหากนำข้อมูลนักเรียนมาเฉลี่ยในแต่ละโรงเรียน แล้ววิเคราะห์ในระดับโรงเรียนจะเรียกว่า การวิเคราะห์แบบรวมกลุ่ม (Aggregated Analysis)
  - เช่น Pornprasertmanit, Lee, & Preacher (2014) พบว่าผล CFA ที่เกิดจากวิเคราะห์ข้อมูลแบบ Disaggreated Analysis จะทำให้น้ำหนักองค์ประกอบ สหสัมพันธ์ระหว่างองค์ประกอบระดับนักเรียน ถูกดึงไปในทิศทางน้ำหนักองค์ประกอบ และสหสัมพันธ์ระดับโรงเรียน ซึ่งไม่ถูกต้อง

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- วิธีการวิเคราะห์สำหรับข้อมูลพหุระดับมีหลายรูปแบบ
  - การใช้การวิเคราะห์สมการเชิงโครงสร้างแบบพหุระดับ (Multilevel Structural Equation Modeling; MSEM) ซึ่งเป็นวิธีการที่ครอบคลุมรูปแบบความสัมพันธ์ระหว่างตัวแปรได้มากที่สุด ทั้งระดับที่ 1 (เช่น นักเรียน) และระดับที่ 2 (เช่น โรงเรียน) ซึ่งเกินขอบเขตเนื้อหานี้
  - การนำกลุ่มมาใช้เป็นตัวแปรอิสระ เรียกว่า แนวทางอิทธิพลคงที่ (Fixed Effect Approach) กล่าวคือ นำตัวแปรโรงเรียน มาแปลงเป็นตัวแปรดัมมี่ แล้วนำมาใช้เป็นตัวแปรอิสระภายนอก (Exogenous Inde pendent Variables)
  - การใช้น้ำหนักตัวอย่าง (Sampling Weights) โดยมองว่าโรงเรียนถูกสุ่มจากประชากรของโรงเรียน และนักเรียนถูกสุ่มจากประชากรนักเรียนภายในโรงเรียน จะตั้งน้ำหนักตัวอย่างของหน่วยนักเรียนในโรงเรียนที่ถูกสุ่มออกมาเยอะๆ ให้น้อยลง เนื่องจากข้อมูลหาง่าย และจะเพิ่มน้ำหนักตัวอย่างของหน่วยนักเรียนในโรงเรียนที่ถูกสุ่มออกมาน้อยๆ ให้มากขึ้น เนื่องจากข้อมูลหายาก ให้น้ำหนักมากกว่า

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- การใช้น้ำหนักกลุ่มตัวอย่าง ให้ K เป็นจำนวนกลุ่ม และ K_P เป็นจำนวนกลุ่มในประชากร n_k เป็นจำนวนตัวอย่างที่สุ่มมาจากกลุ่มที่ k และ N_k เป็นจำนวนประชากรของกลุ่มที่ k โอกาสที่ตัวอย่างที่ i ในกลุ่มที่ k แต่ละหน่วยถูกสุ่มออกมาเป็นดังนี้

$$p_{ik} = \frac{K}{K_P} \cdot \frac{n_k}{N_k}$$

- น้ำหนักของตัวอย่าง (Sampling Weights) คือ ค่าที่บอกว่าหน่วยนั้นเป็นตัวแทนของคนที่คนในประชากร

$$w_{ik} = \frac{1}{p_{ik}}$$

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- สมมติว่าข้อมูล 2 กลุ่มถูกสุ่มมาจากประชากร 10 กลุ่ม และกลุ่มทั้ง 2 กลุ่มมีประชากร 200 และ 300 คนตามลำดับ

ID	*k*	FP1	FP2	*p*	*w*
1	1	10	200	(1/10)(3/200) = 0.0015	666.67
2	1	10	200	(1/10)(3/200) = 0.0015	666.67
3	1	10	200	(1/10)(3/200) = 0.0015	666.67
4	2	10	300	(1/10)(2/300) = 0.00067	1500
5	2	10	300	(1/10)(2/300) = 0.00067	1500

- ค่าน้ำหนักกลุ่มตัวอย่างนี้ จะถูกนำไปใส่ใน sampling.weights ใน lavaan เพื่อให้น้ำหนักแต่ละหน่วยไม่เท่ากัน และให้ใช้ estimator="mlm" เพื่อคำนวณ *SE* และ χ^2 ให้เหมาะสม

ประเด็น	MSEM	Fixed Effect	Sampling Weight
หลักการอ้างอิง	อ้างอิงตามทฤษฎี (Model-based Inference)	อ้างอิงตามทฤษฎี (Model-based Inference)	อ้างอิงตามกรอบประชากร (Design-based Inference)
การอ้างอิงข้อมูล	ประชากรของกลุ่มทั้งหมด	เฉพาะกลุ่มที่เก็บข้อมูลมา	ประชากรของกลุ่มทั้งหมด
ตัวแปรอิสระระดับกลุ่ม	ใช้ได้	ใช้ไม่ได้	ใช้ไม่ได้
การแปลความหมายพารามิเตอร์ระหว่างตัวแปรระดับที่ 1	อิทธิพลภายในกลุ่ม	อิทธิพลภายในกลุ่ม	อิทธิพลไม่สนใจกลุ่ม
จำนวนกลุ่มที่ต้องการ	30 กลุ่มขึ้นไป	2 กลุ่มขึ้นไป	2 กลุ่มขึ้นไป
อิทธิพลที่แตกต่างกันระหว่างกลุ่ม (Cluster-varying Slopes)	อนุญาตตามโมเดลอยู่แล้ว	อนุญาตโดยทำปฏิสัมพันธ์ระหว่างตัวแปรระดับที่ 1 และตัวแปรดัมมี่ของกลุ่ม	ไม่อนุญาต
โมเดล 3 ระดับ	ทำได้	ทำไม่ได้	ทำได้

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- การอ้างอิงตามทฤษฎี (Model-based Inference) จะอ้างอิงไปหาประชากรที่ไม่จำกัด (Infinite population) โดยเชื่อว่าผลที่พบเป็นเรื่องที่เจอโดยทั่วไป ไม่ว่าจะเป็นสถานที่ใด เวลาใด จึงไม่จำเป็นต้องสร้างกรอบประชากรที่ชัดเจน
- การอ้างอิงตามกรอบประชากร (Design-based Inference) จะอ้างอิงไปหาประชากรที่จำกัด มีการนิยามประชากรที่ชัดเจน เช่น โรงเรียนในเขตกรุงเทพมหานคร การสรุปผลจะสรุปเฉพาะกลุ่มคนในประชากรนี้
- อย่างไรก็ตาม การอ้างอิงตามกรอบประชากร อาจมีการอ้างอิงตามทฤษฎีมาเจือปน เช่น อ้างอิงจากเวลาปัจจุบันไปใช้ในอนาคตได้
- ดูรายละเอียดได้ที่ Sterba (2009)

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- การอ้างอิงแบบ Fixed Effect จะอ้างอิงเฉพาะกลุ่มที่เก็บข้อมูลมา เช่น เก็บข้อมูลจาก 5 ประเทศ การอ้างอิงผล จะได้เฉพาะ 5 ประเทศที่เก็บข้อมูลมาเท่านั้น
- การอ้างอิงแบบ Random Effect จะคำนวณไปแล้วว่า 5 ประเทศนี้ถูกสุ่มมาจากประชากรของประเทศทั้งหมด ดังนั้นการอ้างอิงจะสรุปผลไปยังประเทศทั้งหมดได้ ถ้า 5 ประเทศนี้ถูกสุ่มออกมา
- ตัวแปรอิสระระดับกลุ่ม เช่น ทวีปของแต่ละประเทศ ประเภทของโรงเรียนว่าเป็นรัฐบาลหรือเอกชน จะใช้ได้เฉพาะ MSEM

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- สมมติพบผลว่า ระยะทางระหว่างบ้านและโรงเรียนยิ่งสูง ยิ่งมาสายน้อยลง
  - ถ้าเป็น MSEM หรือ Fixed Effect จะเปรียบเทียบระหว่างนักเรียนในโรงเรียนเดียวกัน ถ้านักเรียนคนหนึ่งบ้านอยู่ไกลมากกว่า จะมีโอกาสในการมาสายน้อยกว่านักเรียนอีกคนที่อยู่ใกล้กว่า
  - แต่ถ้าเป็น Sampling Weights คือนำโรงเรียนทั้งหมด ออกจากความสนใจเลย กล่าวคือ นักเรียนคนหนึ่งไม่ว่าจะอยู่โรงเรียนใด ถ้าบ้านอยู่ไกลกว่า จะมีโอกาสในการมาสายน้อยกว่านักเรียนอีกคนที่อยู่ใกล้กว่า ซึ่งอาจอยู่โรงเรียนเดียวกับคนแรกหรือไม่ก็ได้
  - ดูการอภิปรายใน Pornprasertmanit et al. (2014)

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- อิทธิพลที่แตกต่างกันระหว่างกลุ่ม เช่น บางโรงเรียนนักเรียนที่อยู่ใกล้บ้านจะมีอัตราการสายต่ำกว่า แต่ในบางโรงเรียนนักเรียนที่อยู่ใกล้บ้านจะมีอัตราการสายที่มากกว่า ความสัมพันธ์ระหว่างตัวแปรแตกต่างกันระหว่างกลุ่ม
  - MSEM จะอนุญาตให้ทำความแตกต่างระหว่างความสัมพันธ์ได้โดยตรงในโมเดล
  - ใน Fixed Effect สามารถทำปฏิสัมพันธ์ระหว่างกลุ่มและตัวแปรหนึ่งในการทำนายตัวแปรหนึ่ง เพื่อแสดงถึงอิทธิพลที่แตกต่างกันได้
  - ส่วน Sampling Weights จะไปการรวมผลระหว่างโรงเรียน ถือว่าการเกิดขึ้นของปรากฏการณ์ในแต่ละโรงเรียนเหมือนกัน การใส่น้ำหนักเพื่อให้นักเรียนบางคนมีอิทธิพลในการวิเคราะห์มากกว่าบางคนเท่านั้น

2A หน่วยข้อมูลทุกตัวเป็นอิสระจากกัน

- อีกมุมหนึ่ง ที่อาจทำให้ตัวแปรไม่เป็นอิสระจากกัน คือ ในตัวอย่างสามารถแบ่งเป็นกลุ่มย่อยๆ ได้ เช่น ตัวอย่างสามารถแบ่งได้เป็น 2 กลุ่ม คือ กลุ่มที่การบำบัดส่งผลทางบวก และกลุ่มที่การบำบัดไม่ส่งผลใดๆ
- นักวิจัยสามารถใช้โมเดลผสม (Finite Mixture Modeling) ที่สถิติจะวิเคราะห์กลุ่มแฝงที่อยู่ในตัวอย่าง นักวิจัยสามารถกำหนดได้ว่าค่าพารามิเตอร์ใดบ้างที่จะให้แตกต่างกันระหว่างกลุ่มแฝง (เช่น ค่าสัมประสิทธิ์ถดถอย หรือค่าน้ำหนักองค์ประกอบ)
- การวิเคราะห์นี้ค่อนข้างซับซ้อน และปัจจุบันนี้มีเพียง Mplus ที่สามารถวิเคราะห์ Finite Mixture SEM

```
> datsurvey <- read.table("lecture14survey.csv", sep=",", header=TRUE)
> head(datsurvey)
         y1         y2         y3         y4          y5          y6 group fpc1 fpc2 id
1 -0.9905368  1.0450322 -1.2495186 -1.1076858 -0.07036158 -1.03055763     1 1274   40  1
2 -0.3117957  0.3231912  0.2120021  0.8152469 -0.67452245  0.19882280     1 1274   40  2
3  0.7197378 -0.4898080 -0.5378902  1.5078420 -0.42423195  0.14536375     1 1274   40  3
4 -1.6631128 -0.1913558  0.8651971 -0.2528160 -1.08538714 -0.53513779     1 1274   40  4
5  0.2505554 -0.9034588 -1.8965434  0.1613441 -0.05729603  0.89916049     1 1274   40  5
6 -2.7880509  1.2165605 -0.6460378 -1.6815792 -1.95760769 -0.01700837     1 1274   40  6
```

ID ของโรงเรียน

ID ของนักเรียน

จำนวนโรงเรียนในประชากร

จำนวนนักเรียนทั้งหมดในแต่ละโรงเรียน

```
> table(datsurvey$fpc1)

1274
 164
```

นักเรียนทุกคนมีค่าเดียวกัน คือ มีจำนวนโรงเรียนในประชากรทั้งหมด 1,274 โรงเรียน

```
> with(datsurvey, table(group, fpc2))
     fpc2
group 40 50 80 120
    1 40  0  0   0
    2  0  0  0  48
    3  0 27  0   0
    4  0  0  0  19
    5  0  0 30   0
```

ค่าตามจำนวนนักเรียนทั้งหมดในแต่ละโรงเรียน

โรงเรียนที่ 1-5 มี 40, 120, 50, 120, 80 คนตามลำดับ

ซึ่งเก็บข้อมูลมา 40, 48, 27, 19, 30 คนตามลำดับ

วิเคราะห์ CFA แบบปกติ ไม่เอาตัวแปรโรงเรียนไปใส่ในการวิเคราะห์

```
> mdisagg <- '
+ f1 =~ y1 + y2 + y3
+ f2 =~ y4 + y5 + y6
+ '
> outdisagg <- cfa(mdisagg, data=datsurvey)
> summary(outdisagg, fit=TRUE, std=TRUE)
lavaan 0.6-12 ended normally after 36 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        13

  Number of observations                           164

Model Test User Model:

  Test statistic                                 2.344
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.969

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 =~
    y1                1.000                               0.797    0.587
    y2                0.788    0.190    4.153    0.000    0.629    0.522
    y3                0.918    0.214    4.293    0.000    0.732    0.570
  f2 =~
    y4                1.000                               0.883    0.653
    y5                0.923    0.186    4.969    0.000    0.815    0.591
    y6                0.847    0.170    4.994    0.000    0.748    0.599

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 ~~
    f2                0.481    0.125    3.855    0.000    0.683    0.683
```

ค่าน้ำหนักองค์ประกอบมาตรฐานประมาณ .52 - .66

ค่าสหสัมพันธ์ระหว่างองค์ประกอบเท่ากับ .683

Fixed Effect Approach

ใช้ R แปลงให้ตัวแปรดัมมี่ โดยไม่ต้องมาแปลงทีละกลุ่ม

เปลี่ยนตัวแปรกลุ่มให้เป็นรูปแบบ factor ก่อน

```
> datsurvey$groupfac <- factor(datsurvey$group)
> datsurvey$groupfac <- relevel(datsurvey$groupfac, ref="4")
> dummies <- model.matrix(lm(id ~ groupfac, data=datsurvey))
> dummies2 <- cbind(dummies, group=datsurvey$group)
> head(dummies2)
  (Intercept) groupfac1 groupfac2 groupfac3 groupfac5 group
1           1         1         0         0         0     1
2           1         1         0         0         0     1
3           1         1         0         0         0     1
4           1         1         0         0         0     1
5           1         1         0         0         0     1
6           1         1         0         0         0     1
> tail(dummies2)
    (Intercept) groupfac1 groupfac2 groupfac3 groupfac5 group
159           1         0         0         0         1     5
160           1         0         0         0         1     5
161           1         0         0         0         1     5
162           1         0         0         0         1     5
163           1         0         0         0         1     5
164           1         0         0         0         1     5
> colnames(dummies) <- c("intcept", "school1vs4", "school2vs4", "school3vs4", "school5vs4")
> datsurvey <- data.frame(datsurvey, dummies)
```

นำไปแนบกับข้อมูลเดิม

เปลี่ยนกลุ่มอ้างอิง เป็นกลุ่มที่ 4 (อาจไม่เปลี่ยนก็ได้ กลุ่มอ้างอิงจะเป็นกลุ่มแรกเสมอในการแปลงแบบดัมมี่)

แปลงตัวแปรดัมมี่อัตโนมัติ โดยใช้ประโยชน์จากคำสั่ง lm โดยเอาแปรกลุ่มทำนายตัวแปรอะไรก็ได้ เอาผลลัพธ์จากคำสั่ง lm ไปผ่านคำสั่ง model.matrix เพื่อดูว่าตัวแปรอิสระที่แปลงขั้นสุดท้ายก่อนไปทำนายจริงเป็นอย่างไร ซึ่งคำสั่ง lm จะแปลงให้เป็นตัวแปรดัมมี่เรียบร้อย

ทำ dummies2 เพื่อดูว่าแต่ละกลุ่มแปลงเป็นตัวแปรดัมมี่อย่างไร ด้วยคำสั่ง head และ tail

เปลี่ยนชื่อตัวแปรให้เข้าใจว่าเป็นตัวแปรดัมมี่ที่มี 4 เป็นกลุ่มอ้างอิง

```
> mfixed <- '
+ f1 =~ y1 + y2 + y3
+ f2 =~ y4 + y5 + y6
+ y1 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ y2 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ y3 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ y4 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ y5 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ y6 ~ school1vs4 + school2vs4 + school3vs4 + school5vs4
+ '
> outfixed <- cfa(mfixed, data=datsurvey)
> summary(outfixed, fit=TRUE, std=TRUE)
lavaan 0.6-12 ended normally after 81 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        37

  Number of observations                           164

Model Test User Model:

  Test statistic                                 5.139
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.743

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 =~
    y1                1.000                               0.612    0.451
    y2                0.850    0.283    3.001    0.003    0.520    0.432
    y3                1.047    0.343    3.058    0.002    0.641    0.499
  f2 =~
    y4                1.000                               0.739    0.546
    y5                0.995    0.264    3.772    0.000    0.735    0.533
    y6                0.803    0.216    3.709    0.000    0.593    0.475

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 ~~
    f2                0.246    0.090    2.722    0.006    0.544    0.544
```

นำตัวแปรดัมมี่ไปทำนายตัวบ่งชี้ทั้งหมด เพื่อหาผลการทำ CFA เมื่อควบคุมอิทธิพลของกลุ่มทั้งหมด

ถ้าทำนายตัวบ่งชี้ทุกตัว *df* จะไม่เปลี่ยนแปลง กล่าวคือ ไม่ได้ fix parameter อะไรเพิ่มเติม

ค่าน้ำหนักองค์ประกอบมาตรฐานประมาณ .43 - .54

ค่าสหสัมพันธ์ระหว่างองค์ประกอบเท่ากับ .544

Fixed Effect Approach

ใช้ R แปลงให้ตัวแปรดัมมี่ด้วย Effect coding
โดยไม่ต้องมาแปลงทีละกลุ่ม

```
> dummieseff <- model.matrix(lm(id ~ groupfac, data=datsurvey,
+                               contrasts = list(groupfac = contr.sum)))
> dummieseff2 <- cbind(dummieseff, datsurvey$group)
> head(dummieseff2)
  (Intercept) groupfac1 groupfac2 groupfac3 groupfac4
1           1         0         1         0         0 1
2           1         0         1         0         0 1
3           1         0         1         0         0 1
4           1         0         1         0         0 1
5           1         0         1         0         0 1
6           1         0         1         0         0 1
> tail(dummieseff2)
    (Intercept) groupfac1 groupfac2 groupfac3 groupfac4
159           1        -1        -1        -1        -1 5
160           1        -1        -1        -1        -1 5
161           1        -1        -1        -1        -1 5
162           1        -1        -1        -1        -1 5
163           1        -1        -1        -1        -1 5
164           1        -1        -1        -1        -1 5
> colnames(dummieseff) <- c("intcepteff", "school1eff", "school2eff", "school3eff", "school4eff")
> datsurvey <- data.frame(datsurvey, dummieseff)
```

แปลงตัวแปรดัมมี่แบบ Effect coding อัตโนมัติ โดยในคำสั่ง lm ให้กำหนด contrasts ว่า groupfac ให้ใช้ contr.sum ซึ่งเป็นการบอกว่าตัวแปรโรงเรียน ให้ใช้ Effect coding (ซึ่งกลุ่มสุดท้ายจะเป็นกลุ่มอ้างอิง)

ทำ dummieseff2 เพื่อดูว่าแต่ละกลุ่มแปลงเป็นตัวแปรดัมมี่อย่างไร ด้วยคำสั่ง head และ tail

เปลี่ยนชื่อตัวแปรให้เข้าใจว่าเป็นตัวแปรดัมมี่แบบ Effect Coding

นำไปแนบกับข้อมูลเดิม

```
> mfixedeff <- '
+ f1 =~ y1 + y2 + y3
+ f2 =~ y4 + y5 + y6
+ y1 ~ school1eff + school2eff + school3eff + school4eff
+ y2 ~ school1eff + school2eff + school3eff + school4eff
+ y3 ~ school1eff + school2eff + school3eff + school4eff
+ y4 ~ school1eff + school2eff + school3eff + school4eff
+ y5 ~ school1eff + school2eff + school3eff + school4eff
+ y6 ~ school1eff + school2eff + school3eff + school4eff
+ '
> outfixedeff <- cfa(mfixedeff, data=datsurvey)
> summary(outfixedeff, fit=TRUE, std=TRUE)
lavaan 0.6-12 ended normally after 54 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        37

  Number of observations                           164

Model Test User Model:

  Test statistic                                 5.139
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.743

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 =~
    y1                1.000                               0.612    0.451
    y2                0.850    0.283    3.001    0.003    0.520    0.432
    y3                1.047    0.343    3.058    0.002    0.641    0.499
  f2 =~
    y4                1.000                               0.739    0.546
    y5                0.995    0.264    3.772    0.000    0.735    0.533
    y6                0.803    0.216    3.709    0.000    0.593    0.475

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 ~~
    f2                0.246    0.090    2.722    0.006    0.544    0.544
```

นำตัวแปรดัมมี่ไปทำนายตัวบ่งชี้ทั้งหมด เพื่อหาผลการทำ CFA เมื่อควบคุมอิทธิพลของกลุ่มทั้งหมด

วิธีการทำตัวแปรดัมมี่ จะไม่มีผลต่อค่าสหสัมพันธ์ระหว่างตัวบ่งชี้ (เช่น น้ำหนักองค์ประกอบ สหสัมพันธ์ระหว่างองค์ประกอบ) แต่จะไปมีอิทธิพลต่อ meanstructure ดังนั้นถ้าจะใช้ fixed effect approach แล้วมีค่าเฉลี่ย ให้ใช้ Effect coding เพื่อให้ meanstructure หมายถึงค่ากลาง (ค่าเฉลี่ย) จากทุกโรงเรียน

ค่าน้ำหนักองค์ประกอบมาตรฐานประมาณ .43 - .54

ค่าสหสัมพันธ์ระหว่างองค์ประกอบเท่ากับ .544

Sampling Weights

ใส่ตัวแปร ID ของโรงเรียนและนักเรียนตามลำดับ

ใส่จำนวนประชากรระดับโรงเรียนและนักเรียนภายในแต่ละโรงเรียนตามลำดับ

ได้ผลลัพธ์ว่าการสุ่มเป็นแบบสองชั้น

ใช้คำสั่ง weights เพื่อให้ได้น้ำหนักการสุ่ม

นำน้ำหนักที่ได้ไปใส่ใน sampling.weights

อ่านผล Robust standard error

```
> library(survey)
> designdat <- svydesign(id=~group+id, data=datsurvey, fpc=~fpc1+fpc2)
> designdat
2 - level Cluster Sampling design
With (5, 164) clusters.
svydesign(id = ~group + id, data = datsurvey, fpc = ~fpc1 + fpc2)
> datsurvey$w <- weights(designdat)
> mweights <- '
+ f1 =~ y1 + y2 + y3
+ f2 =~ y4 + y5 + y6
+ '
> outweights <- cfa(mweights, data=datsurvey, sampling.weights="w")
> summary(outweights, fit=TRUE, std=TRUE)
lavaan 0.6-12 ended normally after 35 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        13

  Number of observations                           164
  Sampling weights variable                          w

Model Test User Model:
                                              Standard      Robust
  Test Statistic                                 3.108       2.709
  Degrees of freedom                                 8           8
  P-value (Chi-square)                           0.927       0.951
  Scaling correction factor                                  1.147
    Yuan-Bentler correction (Mplus variant)
```

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 =~
    y1                1.000                               0.718    0.545
    y2                0.853    0.212    4.029    0.000    0.612    0.524
    y3                0.958    0.214    4.468    0.000    0.688    0.541
  f2 =~
    y4                1.000                               0.889    0.671
    y5                0.929    0.250    3.712    0.000    0.826    0.598
    y6                0.893    0.200    4.458    0.000    0.793    0.602

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  f1 ~~
    f2                0.452    0.112    4.017    0.000    0.708    0.708
```

ค่าน้ำหนักองค์ประกอบมาตรฐานประมาณ .52 - .68

ค่าสหสัมพันธ์ระหว่างองค์ประกอบเท่ากับ .708

ทำไม Fixed effect approach ให้ค่าสหสัมพันธ์เท่ากับ .544
แต่วิธี Sampling Weights ให้ค่าสหสัมพันธ์เท่ากับ .708

ความหมายของพารามิเตอร์ต่างกัน
วิธีแรกดูสหสัมพันธ์ของค่าภายในโรงเรียน
ส่วนวิธีที่สองดูสหสัมพันธ์ของสองตัวแปรโดยไม่สนใจโรงเรียน

```
> datsurveyg <- aggregate(cbind(y1, y2, y3, y4, y5, y6) ~ group,
+                         data=datsurvey, FUN=mean)
> datsurveyg$f1 <- with(datsurveyg, y1 + y2 + y3)
> datsurveyg$f2 <- with(datsurveyg, y4 + y5 + y6)
> cor(datsurveyg$f1, datsurveyg$f2)
[1] 0.89948
```

ลองใช้ Aggregated approach ประมาณค่าสหสัมพันธ์ระหว่างสองตัวแปรในระดับโรงเรียน พบว่าได้ค่าสหสัมพันธ์เท่ากับ .899

Within-school correlation = .544
Between-school correlation = .899

ผลสหสัมพันธ์โดยไม่สนใจโรงเรียน คือ จุดกึ่งกลางระหว่าง Within- และ Between-school correlation ซึ่งได้ผลจาก Sampling Weights เท่ากับ .708

แม้การแปลความหมายของวิธี Disaggregated และ Sampling weights ใกล้เคียงกัน แต่ Sampling weights ถูกต้องมากกว่า เพราะสนใจจำนวนนักเรียนแต่ละโรงเรียนที่ถูกสุ่มออกมา โดยที่ Disaggregated ไม่สนใจตัวแปรโรงเรียนเลย ให้น้ำหนักนักเรียนทุกคนเท่ากันหมด

```
> mmsem <- '
+ level: 1
+ fw1 =~ y1 + y2 + y3
+ fw2 =~ y4 + y5 + y6
+ level: 2
+ fb1 =~ y1 + y2 + y3
+ fb2 =~ y4 + y5 + y6
+ '
> outmsem <- sem(mmsem, data = datsurvey, cluster = "group")
Warning message:
In lavaan::lavaan(model = mmsem, data = datsurvey, cluster = "group",  :
  lavaan WARNING:
    the optimizer warns that a solution has NOT been found!
```

lavaan ในปัจจุบันสามารถวิเคราะห์ Multilevel SEM แล้ว แต่ในตัวอย่างนี้มีเพียง 5 โรงเรียน ซึ่งไม่เพียงพอในการวิเคราะห์ MSEM ผลจึงออกมาไม่ลู่เข้าหาผลลัพธ์

2B โมเดลเชิงเส้น

- อิทธิพลใน SEM ทั้งหมด ไม่ว่าจะเป็นอิทธิพลระหว่างองค์ประกอบ หรืออิทธิพลจากองค์ประกอบไปหาตัวบ่งชี้ ล้วนเป็นอิทธิพลเชิงเส้นตรง (Linearity)
- การอภิปรายในที่นี้จะแบ่งออกเป็น 2 ส่วน คือ อิทธิพลจากองค์ประกอบไปหาตัวบ่งชี้ และอิทธิพลระหว่างองค์ประกอบกันเอง

2B โมเดลเชิงเส้น

- การตรวจสอบอิทธิพลจากองค์ประกอบไปหาตัวบ่งชี้ ว่าเป็นเส้นตรงหรือไม่ค่อนข้างยาก เพราะคะแนนองค์ประกอบไม่สามารถวัดได้จริง และหากใช้คะแนนองค์ประกอบ (เช่น ผลรวมของตัวบ่งชี้) ก็คือการผูกมัดไปแล้วว่ามีอิทธิพลเชิงเส้นตรง
- ถ้ามีความสัมพันธ์ที่ไม่ใช่เชิงเส้น อาจทำให้ตัวบ่งชี้มีการกระจายไม่เป็นโค้งปกติ ซึ่งการไม่เป็นโค้งปกติอาจเกิดจากได้หลายสาเหตุ ไม่ใช่เฉพาะการละเมิดโมเดลเชิงเส้นอย่างเดียว
- หากตัวบ่งชี้เป็นแบบจัดกลุ่ม แล้วตัวบ่งชี้มีอิทธิพลแบบเส้นโค้งจากองค์ประกอบ โปรแกรมจะปรับจุดเปลี่ยนให้สอดคล้องกับอัตราส่วนของแต่ละกลุ่มในแต่ละตัวบ่งชี้ แล้วอาจทำให้โมเดลเหมาะสมก็ได้ (ทั้งที่มีความสัมพันธ์เชิงเส้นโค้ง)

2B โมเดลเชิงเส้น

- แม้ไม่มีวิธีการเชิงปฏิบัติที่ทำให้ตรวจสอบได้ แต่หากนักวิจัยคาดหวังว่าจะมีความสัมพันธ์จากองค์ประกอบเป็นเส้นโค้ง สามารถใช้ Mplus ประมาณค่าได้ (จากตัวอย่าง Nonlinear CFA)

$$X_{ij} = \nu_i + \lambda_{i1} F_i + \lambda_{i2} F_i^2 + e_{ij}$$

```
TITLE:     this is an example of a non-linear CFA
DATA:      FILE IS ex5.7.dat;
VARIABLE:  NAMES ARE y1-y5;
ANALYSIS:  TYPE = RANDOM;
           ALGORITHM = INTEGRATION;
MODEL:     f BY y1-y5;
           fxf | f XWITH f;
           y1-y5 ON fxf;
OUTPUT:    TECH1 TECH8;
```

2B โมเดลเชิงเส้น

- สำหรับความสัมพันธ์เชิงเส้นโค้งระหว่างองค์ประกอบ สามารถทำ CFA แต่ละองค์ประกอบแยกกัน สร้างคะแนนองค์ประกอบ แล้วนำมาดู Scatterplot เพื่อตรวจสอบว่ามีความสัมพันธ์เชิงเส้นโค้งหรือไม่ (อาจใช้คะแนนรวมแทนได้)
- ในการวิเคราะห์ความสัมพันธ์เชิงเส้นโค้งระหว่างองค์ประกอบ สามารถใช้คำสั่ง XWITH ใน Mplus ได้ (ตัวอย่าง 5.17) (สอดคล้องกับการใช้ nlsem package ใน R) หรือวิธีการสร้างปฏิสัมพันธ์ระหว่างองค์ประกอบวิธีอื่นๆ ซึ่งเกินขอบเขตของวิชานี้

2C ตัวแปรไม่มีความผิดพลาดในการวัด

- วัตถุประสงค์ของการหาความสัมพันธ์ระหว่างองค์ประกอบใน SEM ก็คือต้องการหาความสัมพันธ์ระหว่างตัวแปรโดยตัดอิทธิพลความผิดพลาดในการวัดออก
- อย่างไรก็ตาม SEM ก็อนุญาตให้ใช้ตัวแปรสังเกตได้เสมือนองค์ประกอบในโมเดล เช่น มาใช้เป็นตัวแปรอิสระภายนอก (Exogenous Independent Variable) มาเป็นตัวแปรส่งผ่านหรือมาเป็นตัวแปรตาม
- หากตัวแปรเหล่านั้นมีความผิดพลาดในการวัด ก็ส่งผลต่อขนาดอิทธิพลระหว่างองค์ประกอบ
  - หากเป็นตัวแปรอิสระ (หรือตัวแปรส่งผ่านที่เป็นตัวแปรอิสระของอีกตัวแปร) จะทำให้ค่าสัมประสิทธิ์ถดถอย (*b*) มีขนาดต่ำกว่าปกติ
  - หากเป็นตัวแปรตาม จะทำให้สัมประสิทธิ์การทำนาย (R^2) น้อยกว่าปกติ

2C ตัวแปรไม่มีความผิดพลาดในการวัด

- วิธีการแก้ไข คือ พยายามเก็บตัวบ่งชี้หลายตัวต่อองค์ประกอบ แล้วใช้องค์ประกอบในการหาความสัมพันธ์ หรือหาตัวบ่งชี้ที่เป็นตัวแทนขององค์ประกอบดีมากๆ ความเที่ยงแทบจะสมบูรณ์แบบ
- อย่างไรก็ตาม บางครั้งนักวิจัยมีข้อจำกัด ทำให้ไม่สามารถทำได้ ต้องยอมรับผลกระทบที่เกิดขึ้น

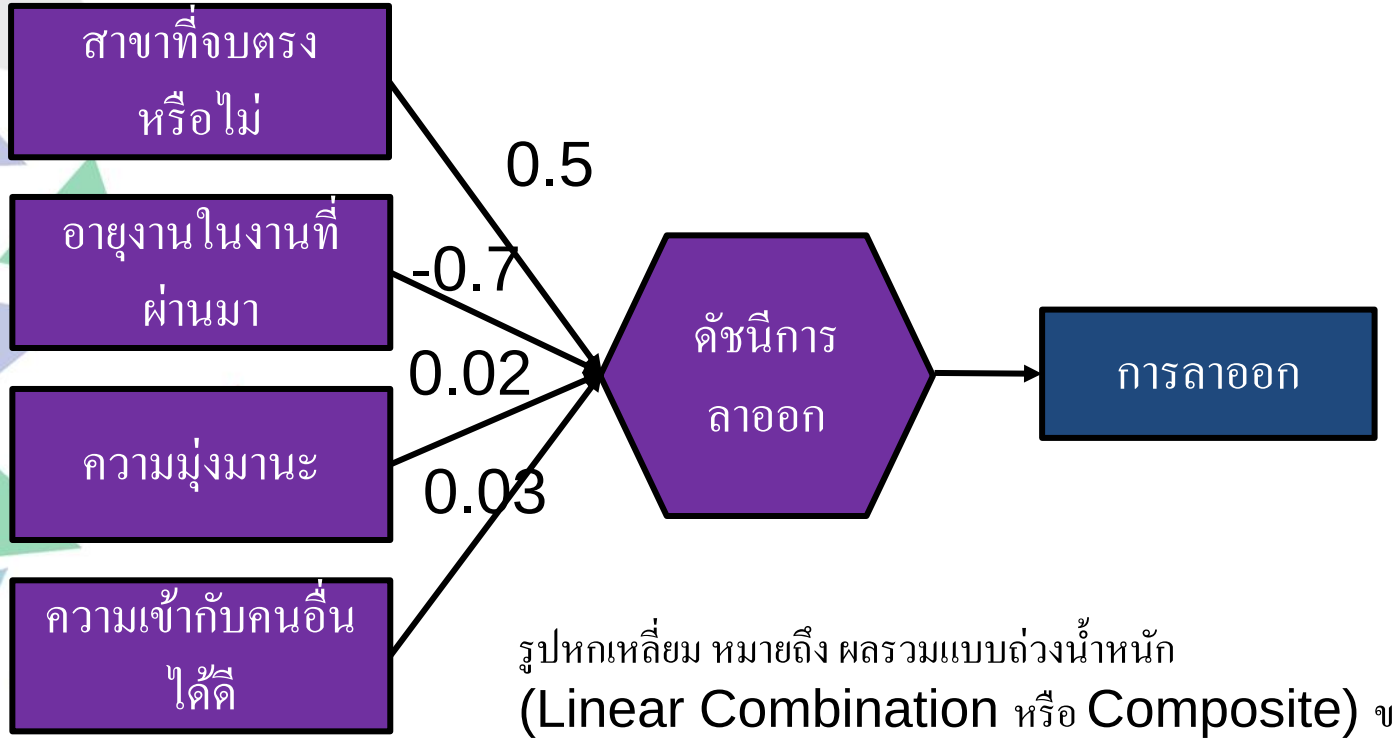
2D องค์ประกอบแบบสะท้อน

- องค์ประกอบแบบสะท้อน (Reflexive Measurement Model) เป็นแนวคิดที่ CFA, EFA รวมถึง SEM ใช้ในการอธิบายความสัมพันธ์ระหว่างตัวบ่งชี้
- มีภาวะสันนิษฐานหนึ่ง ที่เป็นสาเหตุร่วมของตัวบ่งชี้หลายๆ ตัว ส่งผลทำให้ตัวบ่งชี้เหล่านี้มีการเปลี่ยนแปลงพร้อมๆ กัน เช่น ภาวะวิตกกังวล ทำให้ตอบข้อคำถามว่า ฉันตกใจง่าย ฉันกังวลจนนอนไม่หลับ มากขึ้นหรือน้อยลงพร้อมกัน
  - เช่น ความเครียด อาจสะท้อนจาก ความคิดหมกมุ่น อาการนอนไม่หลับ ไม่มีสมาธิ ปวดหัว ฯลฯ อาการเหล่านี้ล้วนมาจากสาเหตุเดียวกัน คือ ความเครียด
  - เราไม่สามารถวัดความเครียดได้โดยตรง แต่ต้องวัดผ่านอาการต่างๆ แล้วนำอาการไปอ้างอิงถึงระดับความเครียด

2D องค์ประกอบแบบสะท้อน

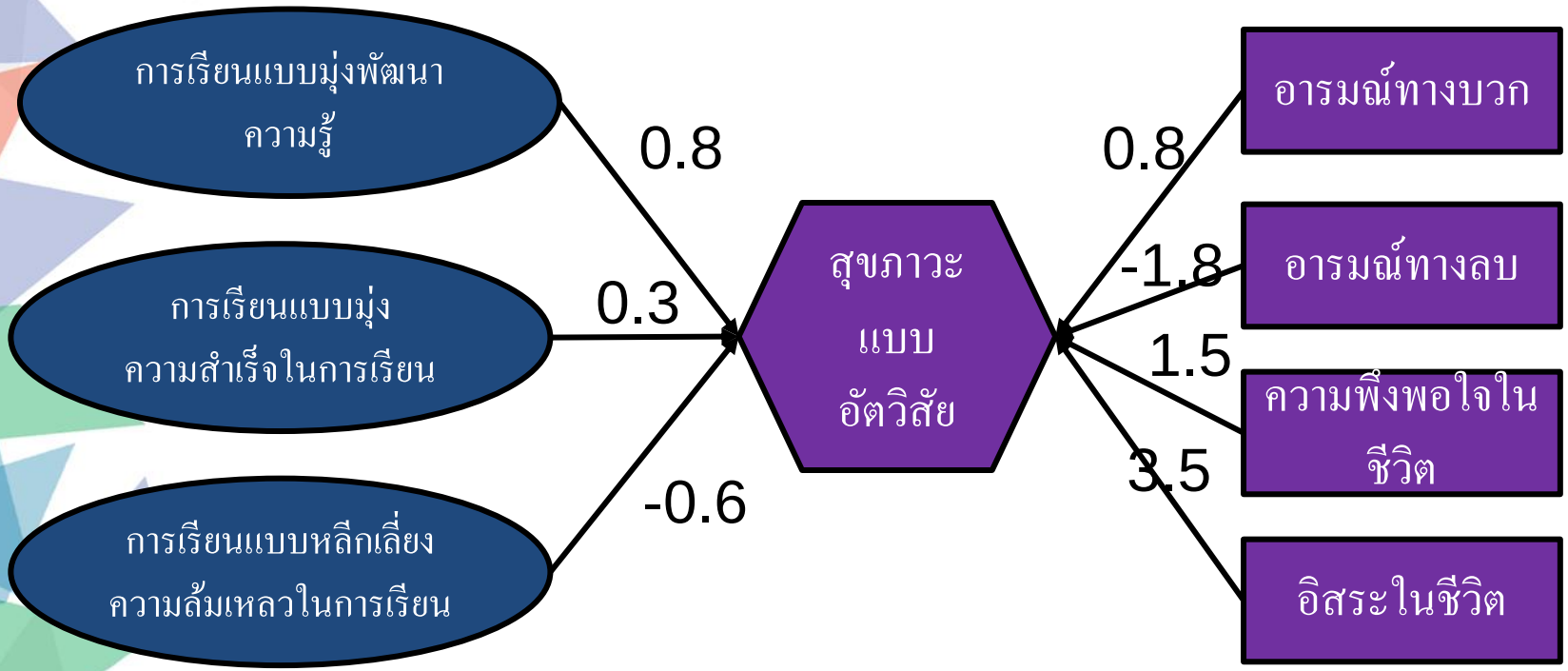
- อย่างไรก็ตาม องค์ประกอบแบบสะท้อนไม่ใช่รูปแบบความสัมพันธ์รูปแบบเดียวในการอธิบายความสัมพันธ์ระหว่างตัวบ่งชี้
- โมเดลการวัดแบบก่อเป็นรูป (Formative Measurement Model) อยู่ในทิศทางตรงข้าม ข้อคำถามหรือตัวบ่งชี้เป็นสาเหตุ ที่ทำให้เกิดผลเป็นภาวะสันนิษฐาน
- เช่น แนวโน้มการลาออก อาจมีตัวบ่งชี้เป็น (ก) สาขาที่เรียนตรงกับตำแหน่งงานหรือไม่ (ข) อายุงานในงานที่ผ่านมา (ค) ความมุ่งมานะ และ (ง) ความเข้ากับคนอื่นได้ดี คะแนนเหล่านี้จะถูกนำมารวมกัน ด้วยการถ่วงน้ำหนักที่แตกต่างกัน เช่น

Quitting Index = 0.5*Area – 0.7*Tenure + 0.02*Grit + 0.03*Agreeableness

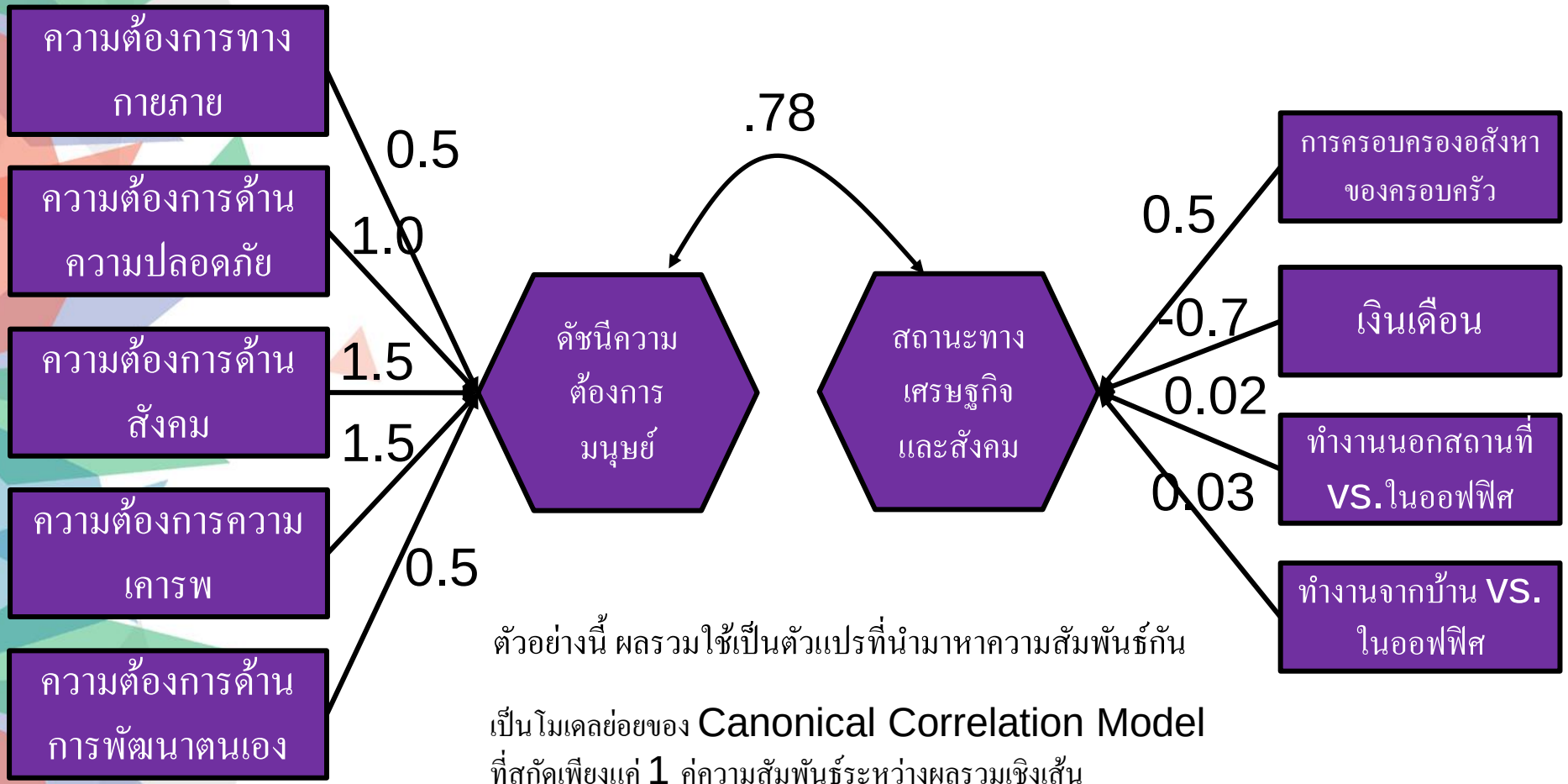


รูปหกเหลี่ยม หมายถึง ผลรวมแบบถ่วงน้ำหนัก
(Linear Combination หรือ Composite) ของตัวแปรต่างๆ

ตัวอย่างนี้ ผลรวมใช้เป็นตัวแปรอิสระ

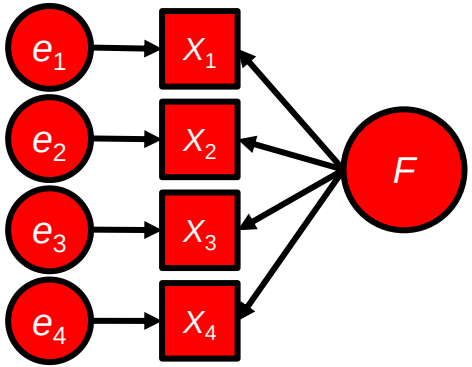


ตัวอย่างนี้ ผลรวมใช้เป็นตัวแปรตาม



ตัวอย่างนี้ ผลรวมใช้เป็นตัวแปรที่นำมาหาความสัมพันธ์กัน

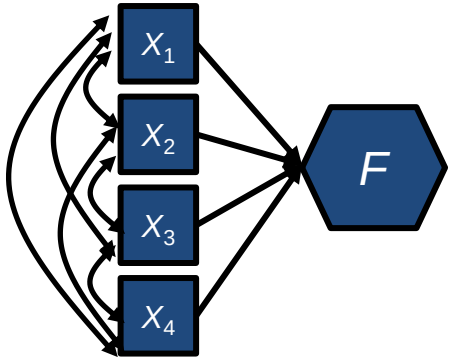
เป็นโมเดลย่อยของ Canonical Correlation Model ที่สกัดเพียงแค่ 1 คู่ความสัมพันธ์ระหว่างผลรวมเชิงเส้น



การวัดแบบสะท้อน (Reflective Measurement Model)

โมเดลองค์ประกอบร่วม (Common Factor Model)

ข้อคำถาม หรือตัวบ่งชี้แต่ละตัว จะสอดคล้องกัน เป็น “ผล” ของสาเหตุเดียวกัน แม้ว่าบางครั้งข้อคำถามจะจับเป็นกลุ่มย่อย ที่เรียกว่าองค์ประกอบ ข้อคำถามในกลุ่มย่อยก็จะเป็น “ผล” ขององค์ประกอบย่อยเดียวกัน



การวัดแบบก่อเป็นรูป (Formative Measurement Model)

โมเดลส่วนประกอบ (Composite Model)

ข้อคำถาม หรือตัวบ่งชี้แต่ละตัว จะเป็นอิสระจากกัน การนำมารวมกัน เหมือนเป็นการสร้างผลรวมเชิงเส้น (Linear Combination) ที่ผลรวมเชิงเส้นนี้ จะถูกนำไปใช้หาความสัมพันธ์กับตัวแปรอื่นให้มากที่สุด ตัวบ่งชี้ไม่จำเป็นต้องสัมพันธ์กันสูง

ประเด็น	สะท้อน	ก่อเป็นรูป
คะแนน	มาตร (Scale)	ดัชนี (Index)
มุ่งหมาย	สะท้อนระดับของภาวะสันนิษฐาน	มัดกลุ่มตัวแปรไว้ด้วยกัน เพื่อความสะดวกในการอธิบายความสัมพันธ์กับตัวแปรอื่นๆ เช่น ดัชนีประสิทธิภาพทีม ที่มีความสัมพันธ์กับรูปแบบภาวะผู้นำ
ความสัมพันธ์กับตัวบ่งชี้	ตัวบ่งชี้เป็นผลที่เกิดขึ้นจากภาวะสันนิษฐาน	ตัวบ่งชี้เป็นสาเหตุที่ทำให้เกิด (ประกอบกันเป็น) ดัชนี
ความสัมพันธ์ระหว่างตัวบ่งชี้	ตัวบ่งชี้ต้องมีความสัมพันธ์กันเองสูง เพราะมีสาเหตุมาจากภาวะสันนิษฐานเดียวกัน	ตัวบ่งชี้ไม่จำเป็นต้องสัมพันธ์กัน ตัวบ่งชี้ที่ดีต้องไม่สัมพันธ์กับตัวบ่งชี้อื่นภายในดัชนี และสัมพันธ์กับตัวแปรนอกดัชนีสูง
การนำตัวบ่งชี้ออก	ความหมายของภาวะสันนิษฐานไม่เปลี่ยนแปลง	ความหมายของดัชนีเปลี่ยน ความสัมพันธ์กับตัวแปรอื่นเปลี่ยน
ลักษณะของตัวบ่งชี้	สะท้อนสิ่งที่ซ่อนอยู่ ไม่สามารถวัดได้โดยตรง มีบางสิ่งบางอย่างบงการตัวบ่งชี้ ที่ทำให้ตัวบ่งชี้มีค่าสูงหรือต่ำ	จะผสมไปด้วยตัวบ่งชี้ ที่ดูแล้วไม่ได้เกี่ยวข้องกับตัวบ่งชี้อื่นเลย อาจเป็นลักษณะของสิ่งแวดล้อมภายนอกด้วยซ้ำ เช่น นำตัวแปรข้อมูลพื้นฐาน (เช่น เพศ อายุ เรียนจบเมืองนอกหรือไม่) มาร่วมกันสร้างดัชนีภาวะผู้นำ
ถ้านำไปวิเคราะห์องค์ประกอบ	ได้การจับกลุ่มของตัวบ่งชี้ชัดเจน	ได้ผลออกมาจับกลุ่มกันได้ไม่หมด ไม่ชัดเจน ตัวแปรสำคัญอาจไม่เข้าพวกกับผู้อื่น
การพัฒนา	ไม่จำเป็นต้องมีตัวแปรอื่นนอกองค์ประกอบมาอ้างอิง เน้นความสัมพันธ์ภายในตัวบ่งชี้	“ควร” ตัวแปรอื่นมาอ้างอิง ความหมายดัชนีขึ้นอยู่กับความสัมพันธ์กับตัวแปรอื่นข้างนอก เช่น นำภาวะผู้นำปฏิรูป แบบแลกเปลี่ยน แบบปล่อยอิสระ มาเป็นตัวแปรตาม เพื่อหาดัชนีภาวะผู้นำจากตัวแปรข้อมูลพื้นฐาน

2D องค์ประกอบแบบสะท้อน

- การใช้โมเดลองค์ประกอบร่วม ไปวิเคราะห์ตัวบ่งชี้ที่ธรรมชาติเป็นโมเดลส่วนประกอบจะมีปัญหาดังต่อไปนี้ (Rhemtulla, van Bork, & Borsboom, 2020)
  - ค่าพารามิเตอร์ระดับโครงสร้างสูงกว่าความเป็นจริง เพราะดึงเฉพาะส่วนความแปรปรวนปรวนที่ร่วมกันมาวิเคราะห์ ทั้งที่โมเดลส่วนประกอบ สนใจใช้ทั้ง common variance และ specific variance ในการหาความสัมพันธ์กับตัวแปรอื่นนอกดัชนี
  - ยิ่งตัวบ่งชี้ในส่วนประกอบมีความสัมพันธ์กันน้อย ค่าพารามิเตอร์ระดับโครงสร้างยิ่งสูงกว่าความเป็นจริงมาก
  - ค่าดัชนีความเหมาะสมของโมเดลไม่ดี แต่บางครั้งโมเดลองค์ประกอบมาวิเคราะห์ส่วนประกอบ อาจเกิดดัชนีความเหมาะสมที่ดีมากได้ แต่ค่าพารามิเตอร์ระดับโครงสร้างจะมีอคติ (bias) สูง
- ดังนั้น นักวิจัยจึงต้องมั่นใจว่าโมเดลการวัดแบบสะท้อนหรือการวัดแบบก่อรูปเหมาะสมกับความสัมพันธ์ระหว่างตัวแปรต่างๆ ของคุณ

2D องค์ประกอบแบบสะท้อน

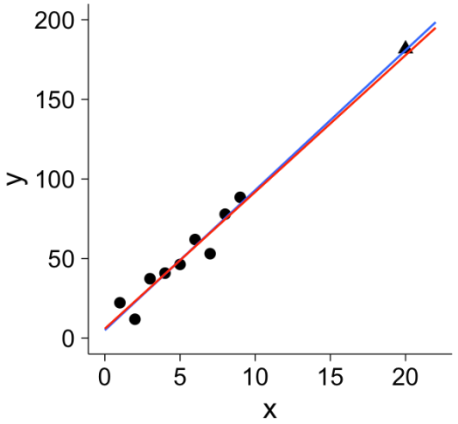
- ลองนึกภาพว่าตัวบ่งชี้ในแต่ละมาตรมีสาเหตุร่วมกันหรือไม่ ถ้าดึงตัวบ่งชี้หนึ่งออกไป ทำให้ความหมายเปลี่ยนหรือไม่ ก่อนจะเลือกใช้การวิเคราะห์องค์ประกอบ
- มาตรที่ดูแล้วอาจเหมาะสมกับโมเดลแบบก่อเป็นรูป เช่น
  - ต้นทุนทางจิตวิทยา (Psychological Capital) คือ ภาวะจิตวิทยาทางบวกที่ส่งผลให้เกิดผลงานที่ดีและประสบความสำเร็จ (แบ่งเป็น Hope, Self-Efficacy, Resilience, Optimism)
  - ผลการปฏิบัติงาน (Job Performance) งานอาจแบ่งออกเป็นส่วนย่อยๆ ที่คะแนนไม่ได้สอดคล้องกัน แต่นักวิจัยหรือองค์กรจำเป็นต้องรวมเพื่อเปรียบเทียบผลงานระหว่างบุคคล
  - คุณภาพของชีวิต (Quality of Life)
  - ภาวะสุขภาวะจิตทางบวก (Subjective Well-being)
  - ความเชี่ยวชาญด้านเทคโนโลยี

2D องค์ประกอบแบบสะท้อน

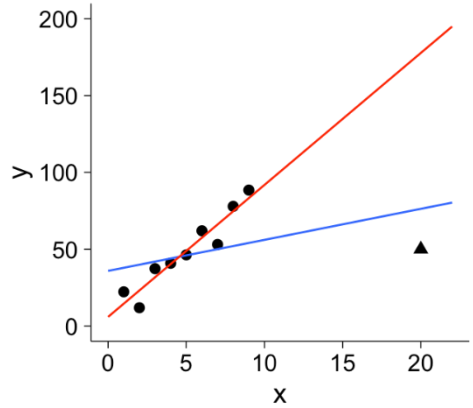
- โมเดลการวัดแบบก่อเป็นรูปใน SEM สามารถวิเคราะห์ด้วย Partial Least Square (PLS) มานานแล้ว ดูรายละเอียดจากหนังสือแนะนำ PLS-SEM โดย Hair et al. (2022)
- เดิมโมเดลการวัดแบบก่อเป็นรูปจะวิเคราะห์ด้วย PLS และโมเดลองค์ประกอบจะใช้โปรแกรม SEM ปกติ ทำให้นักจิตวิทยาส่วนใหญ่ไม่สนใจไปวิเคราะห์ด้วย PLS มากนัก
- ในปัจจุบัน มีวิธีการกำหนดโมเดลส่วนประกอบใน SEM
  - วิธีการที่ดีที่สุดในปัจจุบันให้ดูที่ Yu, Schubert, & Henseler (2023)
  - ดูประวัติการพัฒนาจาก Schubert (2021), Henseler (2021), Gu, Yung, & Cheung (2019), Ogasawara (2007)

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- ผลการวิเคราะห์ที่ได้มา ต้องถูกกำหนดโดยกลุ่มตัวอย่างส่วนใหญ่ ไม่ใช่ผลจากตัวอย่างเพียงไม่กี่กรณี ดังนั้นนักวิจัยควรตรวจสอบว่ามีตัวอย่างใดที่ไปกำหนดทิศทางของการวิเคราะห์มากเกินไปหรือไม่
- ค่าสุดโต่ง (Outliers) คือ กรณีที่ข้อมูลแตกต่างจากข้อมูลอื่น
- กรณีที่มีอิทธิพล (Influential Cases) คือ กรณีที่เปลี่ยนแปลงผลการวิเคราะห์
- ค่าสุดโต่งและกรณีที่มีอิทธิพลไม่ใช่เรื่องเดียวกัน แตกต่างกัน เช่น



ค่าสุดโต่ง แต่ไม่เปลี่ยนสมการถดถอย



ค่าสุดโต่ง และเปลี่ยนสมการถดถอย

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- Pek และ MacCallum (2011) ได้นำแนวคิดการทดสอบค่าสุดโต่งและค่าที่มีอิทธิพลสูงจากการวิเคราะห์ถดถอยมาใช้ใน SEM
- ค่าสุดโต่ง (Outliers) สามารถวิเคราะห์ได้จาก Mahalanobis D^2

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

- ค่านี้จะแสดงความสุดโต่งของแต่ละหน่วยตัวอย่าง ยิ่งค่ามาก ยิ่งสุดโต่ง ในกรณีตัวแปรเดียว ค่า D^2 จะเหมือนกับค่า z^2
- ค่าของแต่ละกรณี สามารถตรวจสอบได้ว่า โอกาสจะเจอค่าที่สุดโต่งขนาดนี้ หากการกระจายประชากรเป็น $\mathrm{MVN}(\bar{x}, S)$ มีมากน้อยเพียงใด โดยเปรียบเทียบ D_i^2 กับ Chi-square distribution ที่ *df* = *p*

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- ทำให้คำนวณค่า *p* ของแต่ละกรณีได้ นักวิเคราะห์สามารถกำหนดได้ ว่าถ้าเจอค่า *p* น้อยมากๆ เช่น < .001 ให้พิจารณากรณีนั้นเป็นพิเศษ
- ใน R มีฟังก์ชั่น mahalanobis อยู่แล้ว เพื่อหาความสุดโต่งของข้อมูล เมื่อการกระจายของประชากรเป็น $\mathrm{MVN}(\bar{\mathbf{x}}, \mathbf{S})$
- อย่างไรก็ตาม การนำ $\bar{\mathbf{x}}, \mathbf{S}$ มาใช้เป็นจุดเปรียบเทียบอาจไม่เหมาะสม เพราะค่า $\bar{\mathbf{x}}, \mathbf{S}$ ก็ได้รับอิทธิพลจากค่าสุดโต่ง
- ใน faoutlier package ได้คำนวณ Mahalonbis D^2 ให้ โดยคำนวณค่า $\bar{\mathbf{x}}, \mathbf{S}$ โดยลดอิทธิพลของค่าสุดโต่งไปด้วย ผมแนะนำให้ใช้วิธีนี้

```
> datcon <- read.table("lecture11consci.csv", sep=",", header=TRUE, na.strings="999")
> mcon <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55'
```

คัดเฉพาะตัวแปรที่จะใช้ในงานวิจัย (ต้องเป็นตัวแปรต่อเนื่อง)

```
> datcon2 <- datcon[,paste0("c", c(1, 7, 13, 19, 25, 31, 37, 43, 49, 55))]
> mh <- mahalanobis(datcon2, center=apply(datcon2, 2, mean), cov=cov(datcon2))
> mh <- data.frame(id=datcon$id, mh=mh)
> mh <- mh[order(mh$mh, decreasing=TRUE),]
> mh$p <- pchisq(mh$mh, df=10, lower.tail=FALSE)
> head(mh)
```

คำสั่ง mahalanobis ใส่ค่าเฉลี่ยและความแปรปรวนร่วมของตัวแปรทั้งหมด

หา *p*-value

```
     id       mh            p
405 405 42.51348 6.074422e-06
453 453 39.10344 2.435118e-05
293 293 38.09016 3.660442e-05
408 408 36.62906 6.559300e-05
37   37 35.86273 8.887263e-05
333 333 32.44421 3.375751e-04
```

```
> library(faoutlier)
> set.seed(123321)
> mhrobust <- robustMD(datcon2)
> mhrobust
         mah       p  sig
405 56.58728 0.00000 ****
293 51.87620 0.00000 ****
453 51.83262 0.00000 ****
408 50.56047 0.00000 ****
37  47.40633 0.00000 ****
333 42.91177 0.00001 ****
433 41.31112 0.00001 ****
274 35.40521 0.00011  ***
317 34.67237 0.00014  ***
391 34.53309 0.00015  ***
```

หา mahalanobis เทียบกับ robust mean และ robust covariance แต่กระบวนการหา robust mean/covariance จะเกี่ยวข้องกับการสุ่ม cases ต่างๆ ออกมา ซึ่งทำให้วิเคราะห์แต่ละครั้ง อาจได้ค่าออกมาแตกต่างกัน ดังนั้น set.seed เป็นการกำหนดให้คอมพิวเตอร์สุ่มจากเลขสุ่มเหล่านี้ เพื่อให้ผลออกมาเหมือนกันทุกครั้งที่ใช้คำสั่งนี้

ค่อนข้างชัด ว่า Case ที่ 405, 293, 453, 408, 37, 333 เป็นค่าสุดโต่ง

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- ค่าที่มีอิทธิพลสูง สามารถมองได้ทั้งความเหมาะสมของโมเดล และค่าพารามิเตอร์
- ค่าที่มีอิทธิพลสูงต่อความเหมาะสมของโมเดล สามารถใช้การเปลี่ยนแปลงของ χ^2 เมื่อกรณีใดถูกตัดออก

$$\Delta\chi_i^2 = \chi_{(i)}^2 - \chi^2 = (N-2)F_{ML(i)} - (N-1)F_{ML} = (N-2)\left(F_{ML} - F_{ML(i)}\right) - F_{ML}$$

- $\chi_{(i)}^2$ และ $F_{ML(i)}$ คือค่าแต่ละค่าที่คำนวณได้เมื่อนำกรณีที่ *i* ออก
- $\Delta\chi_i^2$ มีค่าได้ทั้งค่าบวก ซึ่งหมายความว่านำออกแล้ว ค่าความเหมาะสมดีขึ้น และมีค่าลบ ซึ่งหมายความว่านำออกแล้ว ค่าความเหมาะสมแย่ลง
  - กรณีที่มีค่า $\Delta\chi_i^2$ สูงมากๆ ควรพิจารณาเป็นพิเศษ

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- ค่าที่มีอิทธิพลสูงต่อการประมาณค่าพารามิเตอร์ สามารถใช้ Generalized Cook Distanc e (*gCD*) ในการตรวจสอบอิทธิพลของแต่ละกรณี

$$gCD_i = \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right)' \mathbf{W}_{(i)}^{-1} \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right)$$

- $\hat{\boldsymbol{\theta}}$ คือ เวคเตอร์ของค่าพารามิเตอร์แต่ละตัวมาเรียงกัน อาจใช้ค่าพารามิเตอร์ทุกตัวเลยก็ได้ หรือมุ่งสนใจค่าพารามิเตอร์เพียงแค่บางตัว เช่น ดูเฉพาะน้ำหนักองค์ประกอบเท่านั้น
- $\mathbf{W}$ คือ เมทริกซ์ความแปรปรวนร่วมของค่าสถิติ (Asymptotic Covariance Matrix) ที่รากที่สองของสมาชิกแนวทแยงมุมคือ *SE*
- ค่า *gCD* จะมีค่าต่ำสุดคือ 0 คือเอากรณีนี้ออก ไม่เปลี่ยนค่าพารามิเตอร์เลย และค่า *gCD* เป็นบวก แสดงว่าเกิดการเปลี่ยนแปลง ค่าพารามิเตอร์ทีเปลี่ยนอาจไปทางบวกหรือลบก็ได้

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- แต่อย่าลืมว่า ค่า $\Delta\chi_i^2$ และ gCD_i จะเปลี่ยนแปลงตามโมเดล ถ้าปรับโมเดล ค่านี้ก็จะเปลี่ยนไป
- ผมแนะนำให้พิจารณา $\Delta\chi_i^2$ และ gCD_i ตั้งแต่ช่วงเริ่มต้นในการวิเคราะห์โมเดล เพื่อให้ผลการวิเคราะห์นำนักวิเคราะห์ไปสู่ทิศทางที่ถูกต้อง
  - ถ้าวิเคราะห์องค์ประกอบทีละตัวแปร ให้นำ $\Delta\chi_i^2$ และ gCD_i ของแต่ละตัวแปร มาเรียงเป็นตาราง แล้วดูว่ากรณีไหนควรพิจารณาตัดออก
  - หากรวมองค์ประกอบจากหลายตัวแปรในโมเดลเดียวกันแล้ว (อาจทำ Parceling) อาจตรวจสอบอีกรอบหนึ่ง เพื่อดูว่ามีกรณีใดที่ไปมีอิทธิพลต่อความสัมพันธ์ระหว่างองค์ประกอบ
- เมื่อนักวิเคราะห์ได้โมเดลสุดท้ายแล้ว อาจลองพิจารณาอีกครั้งว่าไม่มีกรณีใดที่มีอิทธิพลต่อผลการวิเคราะห์เป็นพิเศษ แต่ส่วนใหญ่มักไม่เจอตัวใหม่ๆ มักถูกตัดไป หรือรับรู้แล้วตั้งแต่ขั้นตอนที่ผ่านมา

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- แล้วควรตัดค่าสุดโต่ง หรือค่าที่มีอิทธิพลสูงหรือไม่?
  - ถ้าเกิดจากความผิดพลาดในการจัดการข้อมูล เช่น คีย์ผิด กลับคะแนนผิด เผลอกดคีย์บอร์ดทับข้อมูล ลืมกำหนดค่าใดเป็นค่าสูญหาย ให้จัดการทันที แก้ไขให้ถูกต้อง ลบข้อมูลนั้น หรือตัดทิ้งทั้งกรณี
  - ถ้าเกิดจากความผิดพลาดของผู้ให้ข้อมูล เช่น มีคะแนนทิ้งดิ่ง ใช้เวลาเร็วผิดปกติ ให้พิจารณาตัดข้อมูลดังกล่าวเช่นกัน เพราะไม่มีคุณภาพตั้งแต่ต้น
  - ถ้าหาสาเหตุไม่ได้ จะตัดในกรณีที่ชัดเจนจริงๆ ว่าเป็นค่าสุดโต่ง มีอิทธิพลต่อการวิเคราะห์หลายๆ โมเดล จนไม่น่าเอามาวิเคราะห์ต่อ ถ้าไม่ชัด แนะนำว่าอย่าเพิ่งไปตัด
- ค่าที่มีอิทธิพลสูงอาจเกิดจากโมเดลที่ผิด หากโมเดลถูกต้อง ค่าที่มีอิทธิพลสูงอาจหายไป ดังนั้นควรเช็คโมเดลให้มั่นใจก่อนว่าเป็นโมเดลที่ถูกต้อง ก่อนจะตัดกรณีใดจากสถิติอิทธิพลต่างๆ
- ในบางครั้ง การวิเคราะห์นี้จะเรียกว่า การทดสอบความไว (Sensitivity Analysis)

2E ค่าสุดโต่งและค่าที่มีอิทธิพลสูง

- เมื่อมีค่าสุดโต่ง หรือมีอิทธิพล ควรเปลี่ยนไปใช้การวิเคราะห์แบบ MLM หรือ MLR เพราะ SE ที่คำนวณจากเทคนิคเหล่านี้จะลดอิทธิพลจากค่าสุดโต่งโดยธรรมชาติ
- การมีค่าสุดโต่งจับเป็นกลุ่มย่อยๆ อาจแสดงให้เห็นว่าเกิดโมเดลผสม (Mixture Model) ที่กลุ่มย่อยในกลุ่มตัวอย่าง มีค่าพารามิเตอร์ที่แตกต่างกัน เช่น
  - กลุ่มหนึ่ง การบำบัดไม่ได้ผล แต่อีกกลุ่มย่อยหนึ่ง การบำบัดกลับได้ผลดี ซึ่งมีจำนวนไม่เยอะ
  - การวิเคราะห์ผล กลับเห็นว่ากลุ่มย่อยที่สอง เป็นค่าสุดโต่ง เป็นค่าที่มีอิทธิพลสูง ทั้งที่โมเดลผสมน่าจะเหมาะในการวิเคราะห์มากกว่า
- กลุ่มตัวอย่างจะมีผลต่อการวิเคราะห์ค่าสุดโต่งและค่าที่มีอิทธิพลสองด้าน (ก) ยิ่งเยอะ ยิ่งทำให้เวลาการคำนวณเยอะขึ้น (ข) เมื่อนำกรณีออกกรณีหนึ่ง ไม่ได้มีผลกระทบกับค่าพารามิเตอร์เท่าไรนัก ถ้ากลุ่มตัวอย่างสูง ซึ่งทำให้ค่าอิทธิพลต่างๆ มีค่าไม่สูง

```
> library(lavaan)
> mcon <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55'
>
> outcongof <- GOF(datcon2, mcon)
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=28s
> outcongof
          GOF
282  5.65246
455 -5.55283
399 -5.07712
462 -5.01384
408  4.78215
107 -0.02677
162 -0.02471
60  -0.02254
147 -0.02254
2    0.01041
249  0.00630
102 -0.00214
```

X

เครื่องหมายบวก คือ นำออกแล้ว χ^2 สูงขึ้น โมเดลเหมาะสมน้อยลง

เครื่องหมายลบ คือ นำออกแล้ว χ^2 ต่ำลง โมเดลเหมาะสมมากขึ้น

Case ที่ 408 เป็นค่าสุดโต่ง แต่นำออกกลับทำให้ chi-square ดีขึ้น เลยอาจคงไว้ก่อน

```
> outcongcd <- gCD(datcon2, mcon)
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=29s
> outcongcd
          gCD
262 0.6414626
293 0.6104283 X
283 0.5347709
405 0.5312811 X
453 0.4294574 X
274 0.4186868
290 0.3735578
186 0.3722504
266 0.2986516
135 0.2658920
```

Case ที่ 405, 293, 453 เป็นค่าสุดโต่งด้วย และมีค่า gCD สูงด้วย เป็นกรณีที่ควรพิจารณาว่าจะเอาออกหรือไม่

2F ความแปรปรวนของค่าคงเหลือเท่ากัน

- ในการวิเคราะห์ถดถอยจะมีข้อตกลงเบื้องต้นว่าความแปรปรวนของค่าคงเหลือเท่ากัน ในทุกๆ คน หรือที่เรียกว่า Homoscedasticity
- ใน SEM ก็เช่นกัน ค่าคงเหลือต้องมีความแปรปรวนเท่ากัน กล่าวคือ
  - $e_{ij} \sim MVN(0, \theta_{jj})$ หรือ $\zeta_{ik} \sim MVN(0, \psi_{kk})$ กล่าวคือ ในทุกๆ กรณีที่ *i* การกระจายของค่าคงเหลือต้องเท่ากัน คือ θ_{jj} หรือ ψ_{kk}
- ในการวิเคราะห์ถดถอย จะใช้ดูกราฟของค่าคงเหลือ เทียบกับตัวแปรต้น (Residual Plot) ว่ามีการกระจายเป็นรูปใบพัดหรือไม่
- ใน SEM ยังไม่มีคำแนะนำที่ชัดเจนในเรื่องนี้ รอผู้พัฒนาวิธีการตรวจสอบเรื่องนี้

2G การจัดการค่าสูญหายถูกต้อง

- ค่าสูญหาย (Missing Data) หมายถึง ข้อมูลบางตัวแปรจากบางกรณีสูญหายไป ไม่มีข้อมูลมาใช้วิเคราะห์ได้
- หากมีข้อมูลสูญหาย ตามปกติจะตัดข้อมูลของคนนั้นออกไปเลย หรือที่เรียกว่า Listwise Deletion แม้ว่าอาจจะมีข้อมูลบางส่วนอยู่
  - การใช้ Listwise deletion อาจทำให้เกิดความผิดพลาดในการประมาณค่าพารามิเตอร์ และค่า *SE* สูงเกินจริง
- ใน SEM มีวิธีการที่นำข้อมูลส่วนที่ยังคงเหลือในแต่ละคน มาใช้วิเคราะห์ผลทั้งหมด ไม่ว่าจะเป็นวิธีการหาความเป็นไปได้สูงสุดโดยตรง (Direct Maximum Likelihood หรือ Full Information Maximum Likelihood) หรือวิธีการแทนค่าแบบหลายชุด (Multiple Imputation)

2G การจัดการค่าสูญหายถูกต้อง

- ใน Full Information Maximum Likelihood จะใช้เฉพาะข้อมูลที่มี ในการหาค่าความเป็นไปได้สูงสุด
  - สมมติ มีข้อมูล 3 ตัวแปร Y_1, Y_2, Y_3 แล้วมีกลุ่มตัวอย่าง 200 คน มีรูปแบบค่าสูญหาย (Missing Patterns) 2 รูปแบบ โดยที่ 100 คนแรกมีเฉพาะค่า Y_1, Y_2 ขาด Y_3 และอีก 100 คนมีทั้ง 3 ตัวแปร ถ้า

$$\log L = \sum_{i=1}^{N} \left[-\frac{1}{2} (\ln(|\mathbf{\Sigma}|) + (\mathbf{x} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + p \ln(2\pi)) \right]$$

  - จะแปลงตามรูปแบบค่าสูญหาย 2 รูปแบบได้ดังนี้

$$\log L = \sum_{i=1}^{100} \log L_{i(Y_1, Y_2)} + \sum_{i=101}^{200} \log L_{i(Y_1, Y_2, Y_3)}$$

$$\log L = \sum_{i=1}^{100}\left[-\frac{1}{2}\left(\ln\left(\begin{vmatrix}\sigma_{11} & \\ \sigma_{21} & \sigma_{22}\end{vmatrix}\right) + \left(\begin{bmatrix}Y_{i1}\\ Y_{i2}\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\right)' \begin{bmatrix}\sigma_{11} & \\ \sigma_{21} & \sigma_{22}\end{bmatrix}^{-1} \left(\begin{bmatrix}Y_{i1}\\ Y_{i2}\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}\right) + 2\ln(2\pi)\right)\right]$$

$$+ \sum_{i=101}^{200}\left[-\frac{1}{2}\left(\ln\left(\begin{vmatrix}\sigma_{11} & & \\ \sigma_{21} & \sigma_{22} & \\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{vmatrix}\right) + \left(\begin{bmatrix}Y_{i1}\\ Y_{i2}\\ Y_{i3}\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right)' \begin{bmatrix}\sigma_{11} & & \\ \sigma_{21} & \sigma_{22} & \\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{bmatrix}^{-1} \left(\begin{bmatrix}Y_{i1}\\ Y_{i2}\\ Y_{i3}\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right) + 3\ln(2\pi)\right)\right]$$

จะเห็นว่าข้อมูล 100 คนแรก จะใช้แค่ 2 ตัวแปร แต่ข้อมูล 100 คนหลังจะใช้ 3 ตัวแปรในการหาสมาชิกของ $\boldsymbol{\mu}$ และ $\boldsymbol{\Sigma}$

ถ้าใช้ Listwise Deletion ข้อมูลส่วน $\sum_{i=1}^{100} \log L_{i(Y_1,Y_2)}$ จะไม่ถูกใช้ในการหาความเป็นไปได้สูงสุดเลย เพราะ Y_3 หายไป

อย่างไรก็ตาม FIML ใช้ได้ในกรณีที่ตัวแปรภายในมีค่าสูญหายและตัวแปรเป็นตัวแปรต่อเนื่องเท่านั้น หากค่าตัวแปรภายนอกสูญหาย หรือตัวแปรเป็นแบบแบ่งกลุ่มจัดอันดับ จะทำให้ FIML ใช้ไม่ได้ ควรใช้ Multiple Imputation มากกว่า ซึ่งละเอียดเกินไปที่จะสอนในวิชานี้

2G การจัดการค่าสูญหายถูกต้อง

- กระบวนการในการเกิดค่าสูญหายจะแบ่งออกเป็น 3 ประเภท คือ
  - การเกิดค่าสูญหาย เกิดจากการสุ่มที่แท้จริง โอกาสเกิดค่าสูญหายไม่ได้เกี่ยวเนื่องกับตัวแปรใดๆ เลย จะเรียกว่า การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random; MCAR)
  - การเกิดค่าสูญหาย เกิดจากค่าตัวแปรอื่นในโมเดล กล่าวคือ โอกาสเกิดค่าสูญหายถูกกำหนดด้วยตัวแปรที่อยู่ในโมเดล เรียกว่า การสูญหายแบบสุ่ม (Missing at Random; MAR) เช่น เพศชายมีโอกาสเกิดค่าสูญหายในการตอบความพึงพอใจในการแต่งงานมากกว่าเพศหญิง ถ้าโอกาสที่ตัวแปรความพึงพอใจในการแต่งงานจะสูญหาย เปลี่ยนแปลงไปตามเพศอย่างเดียว แล้วเพศอยู่ในโมเดล จะเรียกว่า M AR
  - การเกิดค่าสูญหาย เกิดจากค่าตัวแปรอื่นที่ไม่ได้วัด เช่น ความพึงพอใจในการแต่งงานสูญหาย โดยไม่นำเพศเข้าไปในโมเดล หรือเกิดจากขนาดของตัวแปรนั้นเอง เช่น เงินเดือนถ้ายิ่งสูง โอกาสคนไม่ตอบเงินเดือนยิ่งสูง จะเรียกว่า การสูญหายไม่สุ่ม (Missing Not at Random; MNAR)

2G การจัดการค่าสูญหายถูกต้อง

- FIML หรือ MI จะประมาณค่าพารามิเตอร์ได้ถูกต้อง หรือให้ค่า *SE* ถูกต้อง เมื่อกระบวนการเกิดค่าสูญหายเป็น MCAR หรือ MAR
- กล่าวคือ ให้นำตัวแปรที่ทำให้เกิดโอกาสการเกิดค่าสูญหายแตกต่างกันเข้ามาในโมเดลวิเคราะห์ทั้งหมด เพื่อให้เป็น MAR
- มีประเด็นที่ต้องพิจารณา 3 ประเด็น ในการนำตัวแปรที่เปลี่ยนโอกาสเกิดค่าสูญหายมาใช้ใน FIML
  - นักวิจัยไม่มีทางรู้ ว่าตัวแปรที่เปลี่ยนโอกาสมีอะไรบ้าง (และอย่าลืมว่าตัวแปรที่มีค่าสูญหายอาจมีมากกว่า 1 ตัวแปร ซึ่งตัวแปรที่เปลี่ยนโอกาสเกิดค่าสูญหายของแต่ละตัวแปรไม่เหมือนกัน)
  - ตัวแปรที่เปลี่ยนโอกาส อาจมีหลายตัวแปร นักวิจัยอาจเก็บข้อมูลมาไม่หมด ทำให้เป็นกึ่ง MAR และ MNAR
  - ตัวแปรที่เปลี่ยนโอกาสเกิดค่าสูญหาย อาจไม่ได้เป็นตัวแปรที่สนใจในโมเดล จะนำตัวแปรดังกล่าวมาใช้ใน

2G การจัดการค่าสูญหายถูกต้อง

- จากสองปัจจัยแรก ที่นักวิจัยไม่มีวันรู้ว่าตัวแปรที่ทำให้เกิดค่าสูญหายคืออะไร ซึ่งแน่นอนว่านักวิจัยไม่มีทางรู้ว่าตนเองเก็บข้อมูลตัวแปรที่รับโอกาสเกิดค่าสูญหายมาหมดหรือไม่
- ดังนั้น เมื่อมีค่าสูญหาย นักวิจัยมีโอกาสเจอกระบวนการเกิดค่าสูญหายผสมระหว่าง MAR และ MNAR นักวิจัยต้องเพิ่มโอกาสให้การวิเคราะห์ของตนเป็น MAR และลดโอกาสการเป็น MNAR
- วิธีการลดผลกระทบจากค่าสูญหาย มีดังนี้
    1. ลดอัตราการเกิดค่าสูญหายตั้งแต่ต้น เช่น ในการเก็บข้อมูลระยะยาว มักมีการออกจากงานวิจัย (Dro pout) เป็นเรื่องปกติ แก้ไขไม่ให้สูญหายโดยอาจให้ผู้ตอบกรอกอีเมลล์สำรอง มีรางวัลจูงใจในการตอบซ้ำ ทำการสำรวจให้ไม่น่าเบื่อ เพื่อลดอัตราการออกจากงานวิจัย
    2. ถ้าหลีกเลี่ยงไม่ได้ ให้นักวิจัยคิดล่วงหน้า ว่าตัวแปรอะไรบ้าง ทำให้เกิดค่าสูญหาย เช่น ค่าสูญหายในข้อมูลระยะยาว ให้เก็บข้อมูลพื้นฐานตั้งแต่ต้นให้มากที่สุด อาจถามความเต็มใจที่จะตอบในครั้งถัดไป เป็นต้น

2G การจัดการค่าสูญหายถูกต้อง

- วิธีการลดผลกระทบจากค่าสูญหาย มีดังนี้

3. ถ้ารู้ว่าเป็น MNAR จากตัวแปรตัวนั้นเอง ให้เก็บข้อมูลตัวแปรอื่นมาช่วยทำนายค่าสูญหายให้มากที่สุดเท่าที่ทำได้ เช่น คนที่เงินเดือนสูง มักไม่ค่อยตอบเงินเดือน ให้เก็บข้อมูลเรื่องจำนวนบ้าน จำนวนรถ ค่าใช้จ่ายรายเดือน เพื่อเป็นตัวช่วยทำนายเงินเดือนที่สูญหายไป
4. นำตัวแปรที่สำคัญ ไปเก็บไว้ตั้งแต่ต้นของการสำรวจ เพื่อหลีกเลี่ยงข้อมูลสูญหายในตัวแปรนั้น

- ตัวแปรที่ใช้ในการทำนายข้อมูลสูญหาย แต่ไม่ได้อยากนำเข้าในโมเดล จะเรียกว่า ตัวแปรช่วย (Auxiliary Variables) จะนำเข้าไปในโมเดลวิเคราะห์ได้ 2 วิธี
  - นำตัวแปรช่วยทั้งหมด เป็นตัวแปรภายนอก ทำนายตัวบ่งชี้ทั้งหมดในโมเดล เรียกว่า Fixed Factor Approach
  - นำตัวแปรช่วงทั้งหมด ไปหาความสัมพันธ์กับค่าคงเหลือของตัวบ่งชี้ทุกตัวในโมเดล เรียกว่า Saturated-correlates approach

2G การจัดการค่าสูญหายถูกต้อง

- Correlated-residual approach จะดีกว่า Fixed-factor approach เพราะค่าพารามิเตอร์ในโมเดลไม่เปลี่ยนแปลงไป

```
> set.seed(123321)
> datcon <- read.table("lecture11consci.csv", sep=",", header=TRUE, na.strings="999")
> datconcopy <- datcon
> n <- nrow(datconcopy)
> pmissing <- 1/(1 + exp(-(-1 - 1*scale(datconcopy$Age))))
> datconcopy[runif(n) < pmissing, "c49"] <- NA
```

ใส่ค่าสูญหายเข้าไปในตัวแปร c49 โดยให้โอกาสเกิดค่าสูญหายขึ้นอยู่กับอายุ

ใช้คำสั่ง md.pattern ใน mice package เพื่อดูว่ามีรูปแบบค่าสูญหายกี่รูปแบบ

```
> library(mice)
> md.pattern(datconcopy[,setdiff(colnames(datconcopy), "Job")])
    id c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29 c30 c31 c32
274  1  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
105  1  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
61   1  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
35   1  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
     0  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    c33 c34 c35 c36 c37 c38 c39 c40 c41 c42 c43 c44 c45 c46 c47 c48 c50 c51 c52 c53 c54 c55 c56 c57 c58 c59 c60 Gender Age
274   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1      1   1
105   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1      1   1
61    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1      1   1
35    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1      1   1
      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0      0   0
    Education o1 o2 o3 o4 o5 o6 o7 o8 Salary c49
274         1  1  1  1  1  1  1  1  1      1   1   0
105         1  1  1  1  1  1  1  1  1      1   0   1
61          1  1  1  1  1  1  1  1  1      0   1   1
35          1  1  1  1  1  1  1  1  1      0   0   2
            0  0  0  0  0  0  0  0  0     96 140 236
```

มีรูปแบบค่าสูญหาย 4 รูปแบบ โดยเกิดจากตัวแปร c49 และ Salary คอลัมน์ซ้ายสุด คือ จำนวน Cases ที่มีรูปแบบค่าสูญหายแต่ละแบบ

Listwise Deletion

```
> library(lavaan)
> mcon <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55'
> outcon <- cfa(mcon, data=datconcopy)
> summary(outcon, fit=TRUE, std=TRUE)
lavaan 0.6.16 ended normally after 29 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        20

                                                  Used       Total
  Number of observations                           335         475

Model Test User Model:

  Test statistic                               157.899
  Degrees of freedom                                35
  P-value (Chi-square)                           0.000
```

จำนวนกลุ่มตัวอย่างที่ใช้ คือ 335 จาก 475 คน

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  achi =~
    c1                1.000                               0.469    0.643
    c7                0.933    0.101    9.259    0.000    0.438    0.663
    c13               0.926    0.098    9.473    0.000    0.434    0.688
    c19               0.816    0.141    5.802    0.000    0.383    0.375
    c25               0.275    0.132    2.078    0.038    0.129    0.129
    c31               0.790    0.106    7.470    0.000    0.371    0.501
    c37               0.718    0.109    6.603    0.000    0.337    0.434
    c43               0.637    0.116    5.508    0.000    0.299    0.354
    c49              -0.848    0.128   -6.638    0.000   -0.398   -0.436
    c55              -0.869    0.141   -6.158    0.000   -0.408   -0.401
```

ดู c49, c55 เป็นหลัก

c49 *SE* = .128, น้ำหนักองค์ประกอบมาตรฐาน = -.436

c55 *SE* = .141, น้ำหนักองค์ประกอบมาตรฐาน = -.401

กำหนด missing = “ml” เพื่อทำ FIML

FIML without Auxiliary Variables

```
> outconfiml <- cfa(mcon, data=datconcopy, missing="ml")
> summary(outconfiml, fit=TRUE, std=TRUE)
lavaan 0.6.16 ended normally after 50 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        30

  Number of observations                           475
  Number of missing patterns                         2

Model Test User Model:

  Test statistic                               193.856
  Degrees of freedom                                35
  P-value (Chi-square)                           0.000
```

ใช้ข้อมูลที่มีทั้งหมดจาก 475 คน

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  achi =~
    c1                1.000                               0.466    0.645
    c7                0.895    0.086   10.397    0.000    0.417    0.618
    c13               0.861    0.083   10.358    0.000    0.401    0.624
    c19               0.908    0.123    7.404    0.000    0.423    0.425
    c25               0.324    0.115    2.803    0.005    0.151    0.152
    c31               0.851    0.095    8.963    0.000    0.397    0.520
    c37               0.684    0.095    7.166    0.000    0.318    0.410
    c43               0.676    0.104    6.528    0.000    0.315    0.363
    c49              -0.850    0.129   -6.595    0.000   -0.396   -0.434
    c55              -0.964    0.126   -7.620    0.000   -0.449   -0.427
```

ดู c49, c55 เป็นหลัก

c49 *SE* = .129, น้ำหนักองค์ประกอบมาตรฐาน = -.434
c55 *SE* = .126, น้ำหนักองค์ประกอบมาตรฐาน = -.427

จะเห็นว่า *SE* ของ c49 ใกล้เคียงเดิม แต่ของ c55 น้อยลงชัดเจน
น้ำหนักองค์ประกอบมาตรฐานของ c55 น้อยลงประมาณ .02

FIML with Auxiliary Variables (Saturated-correlates approach)

```
> datconcopy$dgender2 <- datconcopy$Gender == 2
> datconcopy$dgender3 <- datconcopy$Gender == 3
> mconaux <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ dgender2 ~~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ dgender3 ~~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ Age ~~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ Education ~~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ Salary ~~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ dgender2 ~~ dgender3 + Age + Education + Salary
+ dgender3 ~~ Age + Education + Salary
+ Age ~~ Education + Salary
+ Education ~~ Salary
+ '
> outconaux <- cfa(mconaux, data=datconcopy, missing="ml")
> summary(outconaux, fit=TRUE, std=TRUE)
lavaan 0.6.16 ended normally after 175 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                       100

  Number of observations                           475
  Number of missing patterns                         4

Model Test User Model:

  Test statistic                               193.254
  Degrees of freedom                                35
  P-value (Chi-square)                           0.000

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  achi =~
    c1                1.000                               0.465    0.644
    c7                0.897    0.086   10.404    0.000    0.417    0.619
    c13               0.862    0.083   10.359    0.000    0.401    0.624
    c19               0.910    0.123    7.410    0.000    0.424    0.425
    c25               0.324    0.116    2.802    0.005    0.151    0.152
    c31               0.853    0.095    8.967    0.000    0.397    0.520
    c37               0.682    0.095    7.154    0.000    0.318    0.409
    c43               0.677    0.104    6.530    0.000    0.315    0.363
    c49              -0.869    0.129   -6.728    0.000   -0.405   -0.443
    c55              -0.966    0.127   -7.626    0.000   -0.449   -0.427
```

ใส่ auxiliary variables ไปหาความสัมพันธ์กับตัวบ่งชี้ทั้งหมด
และให้มีความสัมพันธ์กันเองทั้งหมด

ใช้ข้อมูลที่มีทั้งหมดจาก 475 คน

ดู c49, c55 เป็นหลัก

c49 *SE* = .129, น้ำหนักองค์ประกอบมาตรฐาน = -.443
c55 *SE* = .127, น้ำหนักองค์ประกอบมาตรฐาน = -.427

จะเห็นว่า *SE* ของทั้งสองตัวใกล้เคียงเดิม
น้ำหนักองค์ประกอบมาตรฐานของ c49 น้อยลงประมาณ .01

FIML with Auxiliary Variables (Fixed-factor approach)

```
> mconfixed <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ c1 ~ dgender2 + dgender3 + Age + Education + Salary
+ c7 ~ dgender2 + dgender3 + Age + Education + Salary
+ c13 ~ dgender2 + dgender3 + Age + Education + Salary
+ c19 ~ dgender2 + dgender3 + Age + Education + Salary
+ c25 ~ dgender2 + dgender3 + Age + Education + Salary
+ c31 ~ dgender2 + dgender3 + Age + Education + Salary
+ c37 ~ dgender2 + dgender3 + Age + Education + Salary
+ c43 ~ dgender2 + dgender3 + Age + Education + Salary
+ c49 ~ dgender2 + dgender3 + Age + Education + Salary
+ c55 ~ dgender2 + dgender3 + Age + Education + Salary
+ '
> outconfixed <- cfa(mconfixed, data=datconcopy, missing="ml", fixed.x=FALSE)
> summary(outconfixed, fit=TRUE, std=TRUE)
lavaan 0.6.16 ended normally after 198 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                       100

  Number of observations                           475
  Number of missing patterns                         4

Model Test User Model:

  Test statistic                               184.344
  Degrees of freedom                                35
  P-value (Chi-square)                           0.000
```

ใส่ auxiliary variables ไปทำนายตัวบ่งชี้ทั้งหมด

ใช้ข้อมูลที่มีทั้งหมดจาก 475 คน

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  achi =~
    c1                1.000                               0.462    0.640
    c7                0.914    0.087   10.449    0.000    0.422    0.627
    c13               0.864    0.084   10.317    0.000    0.400    0.621
    c19               0.855    0.120    7.106    0.000    0.395    0.397
    c25               0.273    0.114    2.384    0.017    0.126    0.127
    c31               0.832    0.094    8.821    0.000    0.385    0.504
    c37               0.670    0.095    7.037    0.000    0.310    0.399
    c43               0.652    0.102    6.374    0.000    0.302    0.348
    c49              -0.888    0.129   -6.880    0.000   -0.411   -0.449
    c55              -0.976    0.127   -7.678    0.000   -0.452   -0.429
```

ดู c49, c55 เป็นหลัก

c49 *SE* = .129, น้ำหนักองค์ประกอบมาตรฐาน = -.449
c55 *SE* = .127, น้ำหนักองค์ประกอบมาตรฐาน = -.429

ค่าออกมาใกล้เคียงกับ saturated-correlates approach
แต่ทางทฤษฎี วิธีนี้จะเปลี่ยนความหมายของค่าพารามิเตอร์

2H Multicollinearity

- ตัวแปรต้นมีความสัมพันธ์กันเองสูง ทำให้อิทธิพลจำเพาะของแต่ละตัวแปรมีน้อย และส่งผลให้ *SE* ของสัมประสิทธิ์ถดถอยของตัวแปรดังกล่างมีค่าสูง
- วิธีการตรวจสอบที่ง่ายที่สุด คือ หาเมทริกซ์สหสัมพันธ์ของตัวแปรต้น (อาจอยู่ในระดับตัวบ่งชี้หรือองค์ประกอบ) จากนั้นตรวจสอบว่ามีค่าสหสัมพันธ์สูงหรือไม่ (เช่น สูงกว่า .9)
- วิธีการที่ดีกว่าการตรวจค่าสหสัมพันธ์ คือ การตรวจ tolerance ซึ่งสามารถนำเมทริกซ์สหสัมพันธ์ไปใส่คำสั่ง lm เพื่อดูว่า R^2 เกิน .9 (ซึ่งก็คือ tolerance ต่ำกว่า .1) หรือไม่
- หากเกิดปัญหา Multicollinearity ให้ตัดตัวแปรตัวหนึ่งออก หรือลองดูว่าสามารถรวมเป็นองค์ประกอบเดียวกันได้หรือไม่

```
> msem <- '
+ achi =~ c1 + c7 + c13 + c19 + c25 + c31 + c37 + c43 + c49 + c55
+ caut =~ c2 + c8 + c14 + c20 + c26 + c32 + c38 + c44 + c50 + c56
+ duti =~ c3 + c9 + c15 + c21 + c27 + c33 + c39 + c45 + c51 + c57
+ orde =~ c4 + c10 + c16 + c22 + c28 + c34 + c40 + c46 + c52 + c58
+ disc =~ c5 + c11 + c17 + c23 + c29 + c35 + c41 + c47 + c53 + c59
+ effi =~ c6 + c12 + c18 + c24 + c30 + c36 + c42 + c48 + c54 + c60
+ sat =~ o5 + o6
+ sat ~ achi + caut + duti + orde + disc + effi
+ '
> outsem <- sem(msem, data=datcon)
> inspect(outsem, "cov.lv")
      achi  caut  duti  orde  disc  effi   sat
achi 0.203
caut 0.009 0.004
duti 0.148 0.015 0.245
orde 0.174 0.018 0.144 0.356
disc 0.137 0.013 0.113 0.171 0.144
effi 0.181 0.009 0.129 0.162 0.145 0.185
sat  0.150 0.010 0.104 0.116 0.122 0.169 0.488
```

ใช้คำสั่ง inspect เพื่อหาค่าความแปรปรวนระหว่างองค์ประกอบทั้งหมด

นำตัวแปรต้นแต่ละตัว ทำนายด้วยตัวแปรต้นตัวอื่น
ถ้าค่าสูงกว่า .9 แสดงว่าอาจมีปัญหา Multicollinearity

```
> covlv <- inspect(outsem, "cov.lv")
> library(psych)
> lmCor(achi ~ caut + duti + orde + disc + effi, data=cov2cor(covlv), n.obs=475)$R2
     achi
0.9120372
> lmCor(caut ~ achi + duti + orde + disc + effi, data=cov2cor(covlv), n.obs=475)$R2
     caut
0.4064073
> lmCor(duti ~ achi + caut + orde + disc + effi, data=cov2cor(covlv), n.obs=475)$R2
     duti
0.5256698
> lmCor(orde ~ achi + caut + duti + disc + effi, data=cov2cor(covlv), n.obs=475)$R2
     orde
0.6372332
> lmCor(disc ~ achi + caut + duti + orde + effi, data=cov2cor(covlv), n.obs=475)$R2
     disc
0.8933445
> lmCor(effi ~ achi + caut + duti + orde + disc, data=cov2cor(covlv), n.obs=475)$R2
     effi
0.9441752
```

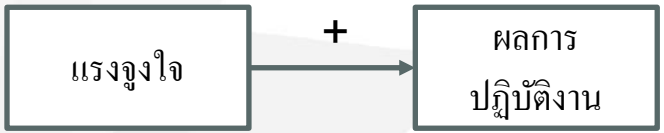
ใช้คำสั่ง lmCor จาก psych package เพื่อทำ lm โดยใส่ Correlation เป็นข้อมูล

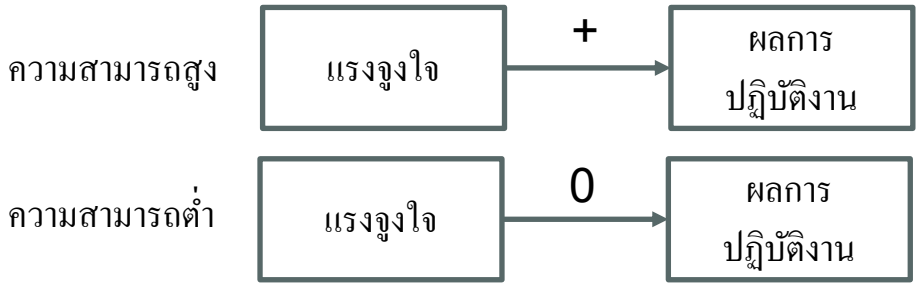
ตัวแปร Achievement Striving และ Self-Efficacy อาจมีปัญหา Multicollinearity

อาจแก้ด้วย Higher-order Factor Model

3 การใส่ตัวแปรที่เกี่ยวข้องทั้งหมด

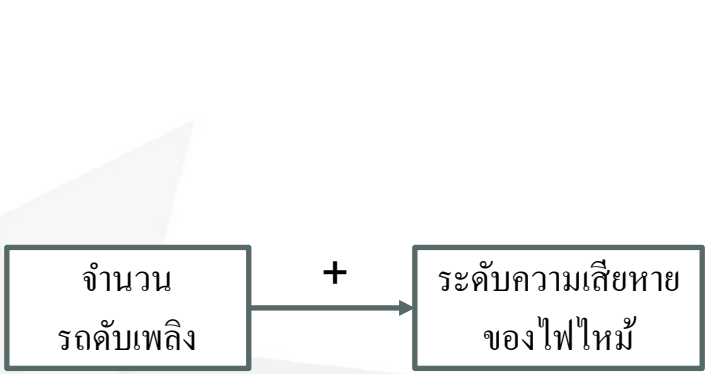
- แม้ว่าโมเดลที่สร้างขึ้นมา จะเหมาะสมอย่างดี แต่โมเดลอาจจะแตกต่างกับความเป็นจริงอย่างมาก ให้ไม่ใส่ตัวแปรที่สำคัญในการอธิบายปรากฏการณ์ทั้งหมด ในที่นี้จะแสดงตัวอย่างไว้ทั้งหมด 3 ตัวอย่าง
- ตัวอย่างที่ 1 การหาอิทธิพลระหว่างตัวแปร โดยไม่ได้ใส่ตัวแปรกำกับที่สำคัญ
  - เช่น แรงจูงใจสัมพันธ์กับผลการปฏิบัติงาน แต่จะได้ผลเฉพาะกับคนที่มีความสามารถ สำหรับคนที่ไม่มีความสามารถ แรงจูงใจที่มากอาจไม่ส่งผล หรือส่งผลทางลบต่อผลการปฏิบัติงาน

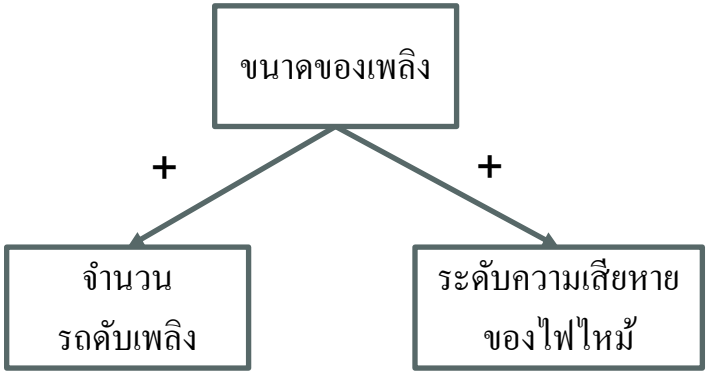




3 การใส่ตัวแปรที่เกี่ยวข้องทั้งหมด

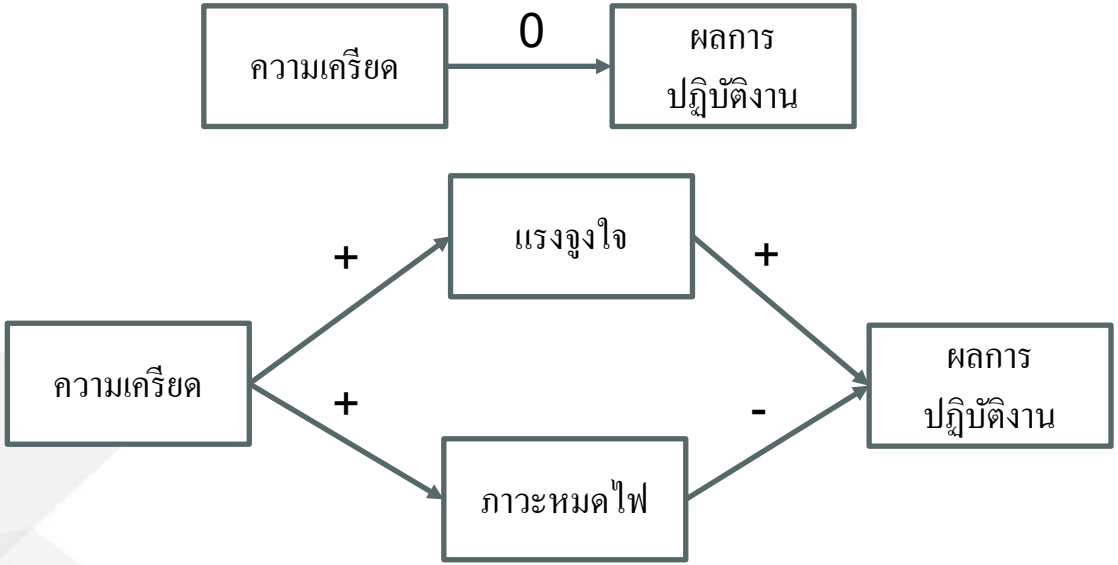
- ตัวอย่างที่ 2 การอธิบายสาเหตุ โดยไม่ได้ใส่สาเหตุร่วมกัน
  - เช่น จำนวนรถดับเพลิง เป็นสาเหตุของความเสียหายจากไฟไหม้ที่เกิดขึ้น ยิ่งจำนวนรถดับเพลิงเยอะ ยิ่งทำให้เกิดความเสียหายที่สูง การวิเคราะห์นี้ลืมใส่ขนาดของไฟ ก่อนที่มีการเรียกรถดับเพลิง





3 การใส่ตัวแปรที่เกี่ยวข้องทั้งหมด

- ตัวอย่างที่ 3 ทิศทางของความสัมพันธ์ โดยไม่ใส่ตัวแปรส่งผ่าน
  - เช่น นักวิจัยอาจพบว่าความเครียด ไม่ส่งผลต่อผลการปฏิบัติงาน แต่พอใส่ตัวแปรส่งผ่านไปสองตัว ตัวหนึ่งคือแรงจูงใจในการทำงาน อีกตัวหนึ่งคือภาวะหมดไฟ



3 การใส่ตัวแปรที่เกี่ยวข้องทั้งหมด

- วิธีแก้ที่ดีที่สุด คือ ทบทวนวรรณกรรมให้รอบคอบ และคิดไตร่ตรองตอนสร้างสมมติฐาน เสมอว่า “มีคำอธิบายอื่นที่เป็นไปได้หรือไม่”
- คำถามสำหรับ Critical Thinking
  - What do you mean?
  - How do you know?
  - Is it true?
  - Can it be explained otherwise?