ทบทวนการวิเคราะห์ถดถอย

โมเดลพหุระดับ (Multilevel Modeling)

สันทัด พรประเสริฐมานิต

โครงร่างการนำเสนอ

- มโนทัศน์พื้นฐาน
- สัมประสิทธิ์ถดถอยมาตรฐาน (Standardized Regression Coefficients)
- การทดสอบสมมติฐาน
- การเขียนรายงาน
- ค่าสัมประสิทธิ์ถดถอยและค่าสัมประสิทธิ์สหสัมพันธ์
- ความสัมพันธ์เชิงเหตุผล (Causal Relationship)
- การเปลี่ยนศูนย์กลาง (Centering)
- ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ
- การวิเคราะห์ถดถอยแบบลำดับขั้น

มโนทัศน์พื้นฐาน

- การวิเคราะห์ถดถอยแบบพหุ (Multiple Regression) เป็นโมเดลที่มีลักษณะดังนี้ สมการเชิงเส้นจะเป็นดังต่อไปนี้

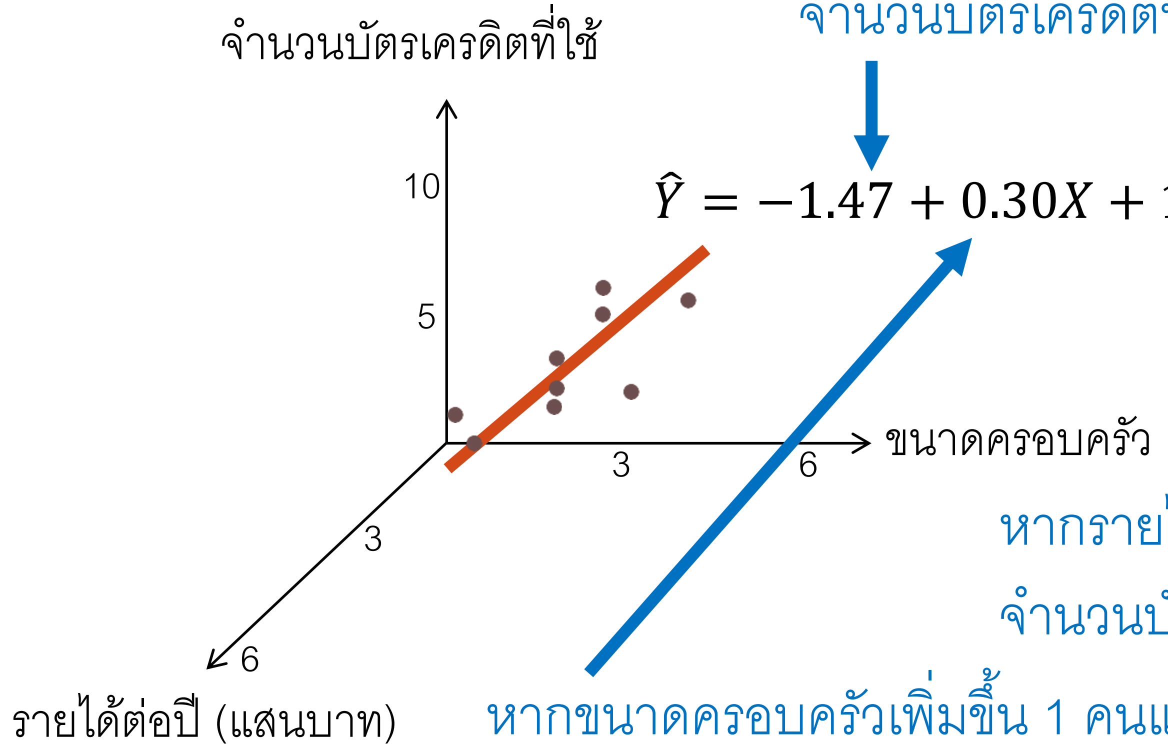
$$Y = \hat{Y} + e = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p + e$$

- a คือ ค่าของ $\hat{Y}$ เมื่อ $X_1, X_2, \ldots, X_p$ เท่ากับ 0 หรือจุดตัดแกน Y ของสมการถดถอย
- b_p คือ ค่าของ $\hat{Y}$ เมื่อ X_p เพิ่มขึ้น 1 หน่วย และควบคุมให้ค่าของ $X_1, X_2, \ldots, X_{p-1}$ คงที่

มโนทัศน์พื้นฐาน

- เช่น ทำนายบัตรเครดิตที่ใช้ ด้วยขนาดครอบครัว และรายได้ครอบครัวต่อปี (หน่วยแสนบาท)

หากขนาดครอบครัวเท่ากับ 0 คนและรายได้ต่อปีเท่ากับ 0 บาท
จำนวนบัตรเครดิตที่ใช้เท่ากับ -1.47 ใบ



$$\hat{Y} = -1.47 + 0.30X + 1.62Z$$

หากรายได้ต่อปีเพิ่มขึ้น 1 แสนบาทและขนาดครอบครัวคงที่
จำนวนบัตรเครดิตที่ใช้จะเพิ่มขึ้น 1.62 ใบ

หากขนาดครอบครัวเพิ่มขึ้น 1 คนและรายได้ต่อปีถูกควบคุมให้คงที่
จำนวนบัตรเครดิตที่ใช้จะเพิ่มขึ้น 0.30 ใบ

มโนทัศน์พื้นฐาน

- สมการ เป็นเพียงเส้นที่บอกว่า ควรแปลงค่า X อย่างไรที่จะทำนายค่า Y ได้ดีที่สุด แต่ไม่ได้บอกความแม่นยำในการทำนาย
- ความแม่นยำในการทำนาย อาจมองได้ว่าเป็นการเกาะกลุ่มของข้อมูลรอบๆ เส้นถดถอย
  - หากข้อมูลเกาะกลุ่มมาก แสดงว่าแม่นยำมาก
- สามารถวัดได้ง่ายๆ ด้วยการหาค่าสหสัมพันธ์ระหว่าง Y และ $\hat{Y}$ จะเรียกค่านี้ว่า ค่าสหสัมพันธ์แบบพหุ (Multiple Correlation; *R*)

มโนทัศน์พื้นฐาน

- ในกรณีที่มีตัวแปรอิสระตัวเดียว ค่า R จะเท่ากับค่าสัมบูรณ์ของค่าสหสัมพันธ์ระหว่าง 2 ตัวแปร, $| r |$
- สัมประสิทธิ์การทำนาย (Coefficient of Determination, R^2) เป็นค่าที่บอกว่าตัวแปร X ทั้งหมดสามารถอธิบายการเปลี่ยนแปลง (ความแปรปรวน) ของตัวแปร Y ได้เท่าไร
  - เช่น จำนวนสมาชิกในครอบครัว สามารถทำนายความแปรปรวนของจำนวนบัตรเครดิตที่ใช้เท่ากับ 75%
  - จำนวนสมาชิกในครอบครัว และรายได้ต่อปี สามารถทำนายความแปรปรวนของจำนวนบัตรเครดิตที่ใช้เท่ากับ 89%

สัมประสิทธิ์การถดถอยมาตรฐาน

- ในการวิเคราะห์ถดถอยแบบพหุ ให้แปลงตัวแปรทุกตัวเป็นคะแนนมาตรฐาน (Standard Score) และสร้างสมการถดถอย

$$\hat{z}_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2} + \cdots + \beta_p z_{X_p}$$

จุดตัดแกน Y เท่ากับศูนย์ เพราะ เส้นสมการถดถอยผ่านจุดค่าเฉลี่ยของตัวแปรทุกตัวเสมอ กล่าวคือ $\hat{z}_Y = 0$ เมื่อ $z_{X_1} = z_{X_2} = \cdots = z_{X_p} = 0$

สัมประสิทธิ์การถดถอยมาตรฐาน

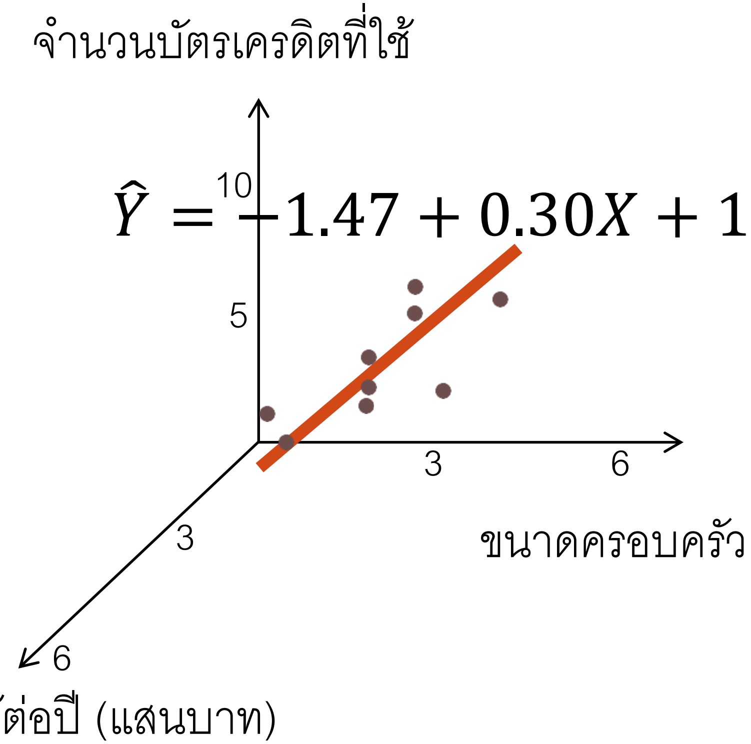
- ค่าสัมประสิทธิ์ถดถอยมาตรฐานนี้ จะมีประโยชน์ในการเปรียบเทียบอำนาจในการทำนายของตัวแปรภายในสมการ
- กล่าวคือ หากค่าสัมบูรณ์ของสัมประสิทธิ์ถดถอยมาตรฐานของตัวแปรใดสูงกว่า แสดงว่าตัวแปรนั้นมีอิทธิพลมากกว่า

สัมประสิทธิ์การถดถอยมาตรฐาน

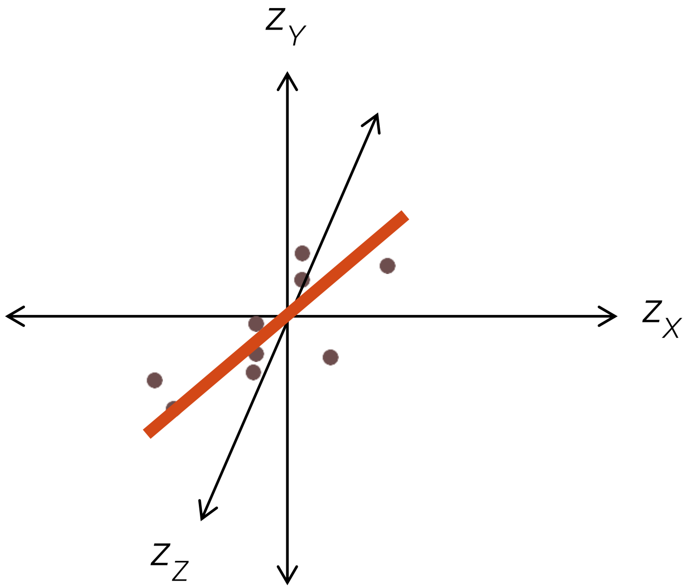
- เช่น ตัวอย่างจำนวนบัตรเครดิต

$R^2 = .89$

$$\hat{Y} = -1.47 + 0.30X + 1.62Z$$



$$\hat{z}_Y = 0.27z_X + 0.71z_Z$$



อิทธิพลของรายได้ต่อปีสูงกว่าขนาดครอบครัวมาก

สัมประสิทธิ์การถดถอยมาตรฐาน

- (ข้อควรระวัง) คนส่วนใหญ่มักเข้าใจว่า ค่าสัมประสิทธิ์ถดถอยมาตรฐาน จะมีค่าระหว่าง -1 ถึง 1 เท่านั้น แท้จริงแล้วไม่จริง
- หากมีตัวแปรอิสระมากกว่า 2 ตัว และความสัมพันธ์ระหว่างตัวแปรอิสระไม่เท่ากับ 0 แล้ว ค่าสัมประสิทธิ์ถดถอยมาตรฐานอาจสูงกว่า 1 หรือต่ำกว่า -1 ได้
- โดยเฉพาะอย่างยิ่ง ความสัมพันธ์ระหว่างตัวแปรอิสระสูงมากๆ

$r(X_1, X_2) = .5$

ในเซลล์แสดงถึง β_1, β_2

$r(Y, X_1) \setminus r(Y, X_2)$	.1	.2	.3	.4	.5	.6
.1	.067, .067	.000, .200	-.067, .333	-.133, .467	-.200, .600	-.267, .733
.2	.200, .000	.133, .133	.067, .267	.000, .400	-.067, .533	-.133, .667
.3	.333, -.067	.267, .067	.200, .200	.133, .333	.067, .467	.000, .600
.4	.467, -.133	.400, .000	.333, .133	.267, .267	.200, .400	.133, .533
.5	.600, -.200	.533, -.067	.467, .067	.400, .200	.333, .333	.267, .467
.6	.733, -.267	.667, -.133	.600, .000	.533, .133	.467, .267	.400, .400

$r(X_1, X_2) = .9$

ในเซลล์แสดงถึง β_1, β_2

$r(Y, X_1) \setminus r(Y, X_2)$	.1	.2	.3	.4	.5	.6
.1	.052, .052	.421, .579	-.895, 1.105	-1.368, 1.631	-1.842, 2.158	
.2	.579, .421	.105, .105	-.368, .632	-.842, 1.158	-1.316, 1.684	-1.789, 2.211
.3	1.105, -.895	.632, -.368	.158, .158	-.316, .684	-.789, 1.211	-1.263, 1.737
.4	1.631, -1.368	1.158, -.842	.684, -.316	.211, .211	-.263, .737	-.737, 1.263
.5	2.158, -1.842	1.684, -1.316	1.211, -.789	.737, -.263	.263, .263	-.211, .789
.6		2.211, -1.789	1.737, -1.263	1.263, -.737	.789, -.211	.316, .316

สัมประสิทธิ์การถดถอยมาตรฐาน

- (ข้อควรระวัง) สัมประสิทธิ์ถดถอยมาตรฐาน สามารถใช้เปรียบเทียบอิทธิพลของตัวแปรในสมการเดียวกันเท่านั้น ไม่สามารถเปรียบเทียบค่ากับสมการอื่นได้

$$\hat{z}_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2} + \cdots + \beta_p z_{X_p}$$

เปรียบเทียบกันได้

เปรียบเทียบกันไม่ได้

กลุ่มที่ 1

$$\hat{z}_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2}$$

กลุ่มที่ 2

$$\hat{z}_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2}$$

สัมประสิทธิ์การถดถอยมาตรฐาน

- (ข้อควรระวัง) คนส่วนใหญ่มักเข้าใจว่า ค่าสัมประสิทธิ์ถดถอยมาตรฐาน และค่าสหสัมพันธ์ต้องเป็นไปในทิศทางเดียวกัน
- จริงๆ แล้วไม่ใช่ ถึงแม้ว่าค่าสหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามในการวิเคราะห์ถดถอยแบบพหุเป็นไปในทางบวก ความสัมประสิทธิ์แบบถดถอยอาจเป็นไปในทางลบ

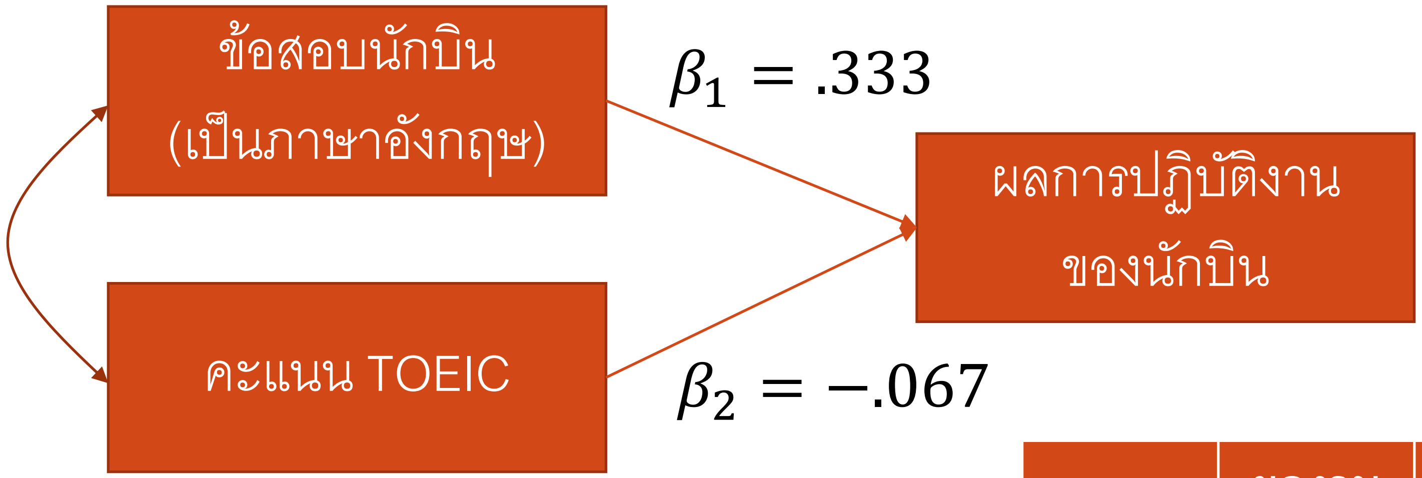
$r(X_1, X_2) = .5$

ในเซลล์แสดงถึง β_1, β_2

$r(Y, X_1) \setminus r(Y, X_2)$	.1	.2	.3	.4	.5	.6
.1	.067, .067	.000, .200	-.067, .333	-.133, .467	-.200, .600	-.267, .733

สัมประสิทธิ์การถดถอยมาตรฐาน

- ตัวอย่างของกรณีนี้ เช่น



	ผลงาน	ข้อสอบ	TOEIC
ผลงาน			
ข้อสอบ	.3		
TOEIC	.1	.5	

การทดสอบสมมติฐาน

- การทดสอบสมมติฐานในการวิเคราะห์ถดถอยแบบพหุ สามารถทำได้ 2 รูปแบบ คือ
  - ค่าพารามิเตอร์ของ R^2 ซึ่งเป็นการทดสอบว่าตัวแปรอิสระทั้งชุด สามารถทำนายตัวแปรตามในประชากรได้อย่างมีนัยสำคัญหรือไม่
  - ค่าพารามิเตอร์ของ a และ b แต่ละตัว ซึ่งเป็นการทดสอบว่าจุดตัดและความชันในประชากรแตกต่างจากศูนย์อย่างมีนัยสำคัญหรือไม่ (ขึ้นอยู่กับความหมายของ a และ b)

การทดสอบสมมติฐาน

นักวิจัยต้องการทดสอบว่าความกดดันจากทางบ้านและความกดดันจากที่ทำงาน อะไรทำนายความเครียดมากกว่ากัน

นักวิจัยจึงเก็บข้อมูลจากพนักงานบริษัทจำนวน 200 คน

สมมติฐาน คือ ความกดดันทั้งสองแบบสามารถทำนายความเครียด

การทดสอบสมมติฐาน

```
> ex1 <- read.table("lecture2ex1.csv", sep=",", header=TRUE)
> library(psych)
> describe(ex1)
          vars   n  mean    sd median trimmed   mad   min   max range  skew kurtosis   se
homepress    1 100 50.25 10.76  51.00   50.76 11.12 18.00 76.00 58.00 -0.44     0.15 1.08
workpress    2 100 51.03 10.63  51.00   51.34  9.64 22.00 73.00 51.00 -0.34     0.02 1.06
stress       3 100 52.10 11.33  53.92   52.12 12.91 22.81 78.09 55.27 -0.07    -0.60 1.13

> corr.test(ex1)
Call:corr.test(x = ex1)
Correlation matrix
          homepress workpress stress
homepress      1.00      0.55   0.57
workpress      0.55      1.00   0.48
stress         0.57      0.48   1.00
Sample Size
[1] 100
Probability values (Entries above the diagonal are adjusted for multiple tests.)
          homepress workpress stress
homepress         0         0      0
workpress         0         0      0
stress            0         0      0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

ความสัมพันธ์ของตัวแปรแต่ละต้นอยู่ระหว่างสูงหรือค่อนข้างสูง และถึงระดับนัยสำคัญ

การทดสอบสมมติฐาน

```
> out1 <- lm(stress ~ homepress + workpress, data=ex1)
> summary(out1)

Call:
lm(formula = stress ~ homepress + workpress, data = ex1)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.2903  -5.4976   0.2526   5.7299  23.3002 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.0114     5.0209   3.189  0.00192 ** 
homepress     0.4588     0.1026   4.473 2.09e-05 ***
workpress     0.2553     0.1038   2.461  0.01564 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.142 on 97 degrees of freedom
Multiple R-squared:  0.3626,	Adjusted R-squared:  0.3495 
F-statistic:  27.6 on 2 and 97 DF,  p-value: 3.254e-10
```

$R^2 = .363$, $F(2, 97) = 27.6$, $p < .001$, Adjusted $R^2 = .350$, $s_e = 9.14$

การทดสอบสมมติฐาน

$$\hat{Y} = 16.011 + 0.459X_1 + 0.255X_2$$

```
> out1 <- lm(stress ~ homepress + workpress, data=ex1)
> su

Call:
lm(formula = stress ~ homepress + workpress, data = ex1)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.2903  -5.4976   0.2526   5.7299  23.3002 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.0114     5.0209   3.189  0.00192 ** 
homepress     0.4588     0.1026   4.473 2.09e-05 ***
workpress     0.2553     0.1038   2.461  0.01564 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.142 on 97 degrees of freedom
Multiple R-squared:  0.3626,	Adjusted R-squared:  0.3495 
F-statistic:  27.6 on 2 and 97 DF,  p-value: 3.254e-10
```

ความกดดันจากทางบ้านและที่ทำงาน
มีผลอย่างมีนัยสำคัญทั้งคู่

การทดสอบสมมติฐาน

```
> library(betaDelta)
> BetaDelta(out1)
Call:
BetaDelta(object = out1)

Standardized regression slopes with MVN standard errors:
             est     se      t df      p   0.05%    0.5%   2.5%  97.5%  99.5% 99.95%
homepress 0.4355 0.0898 4.8509 97 0.0000  0.1308  0.1996 0.2573 0.6137 0.6714 0.7402
workpress 0.2396 0.0944 2.5388 97 0.0127 -0.0807 -0.0084 0.0523 0.4268 0.4875 0.5598
```

อิทธิพลจากทางบ้านสูงกว่าชัดเจน

```
> anova(out1)
Analysis of Variance Table

Response: stress
          Df Sum Sq Mean Sq F value    Pr(>F)    
homepress  1 4106.5  4106.5 49.1369 3.187e-10 ***
workpress  1  506.0   506.0  6.0547   0.01564 *  
Residuals 97 8106.6    83.6
```

ใน R จะไม่มีตาราง ANOVA เหมือนใน SPSS แต่ท่านสามารถคำนวณ $SS_{\text{Regression}}$ โดยนำ *SS* และ *df* ของตัวแปรทั้งสองมารวมกัน แล้วหา *MS* ด้วยตนเอง (ค่า *F* มีในผลของ `summary` แล้ว)

อย่าใช้ตาราง ANOVA ที่แสดงนี้โดยไม่ระวัง เพราะตารางนี้เป็นผลจากการใส่ตัวแปรทีละตัว กล่าวคือ ค่า *p* จะบอกถึงผลของตัวแปรดังกล่าว เมื่อควบคุมตัวแปรที่ใส่ก่อนหน้า ซึ่งแปลความหมายยาก

การทดสอบสมมติฐาน

นักวิจัยต้องการทดสอบว่าความกดดันจากทางบ้านและความกดดันจากที่ทำงาน อะไรมีผลต่อความเครียดมากกว่ากัน

สรุป

ตัวแปรอิสระทั้งสองสามารถอธิบายการกระจายของความเครียดได้ 36%

ทั้งความกดดันจากทางบ้านและความกดดันจากที่ทำงานมีผลอย่างมีนัยสำคัญ แต่ความกดดันจากทางบ้านมีอิทธิพลมากกว่า

การทดสอบสมมติฐาน

ท่านต้องการทดสอบว่าอะไร
ทำให้คนอื่นมองว่าคนน่ารัก
ท่านจึงทำนายความน่ารักด้วย
(1) หน้าตาดี (2) หุ่นดี (3) นิสัยดี
(4) ความใกล้ชิดสนิทสนม

นักวิจัยจึงเก็บข้อมูลจากนิสิตระดับปริญญาตรี
จากวิชาหนึ่ง จำนวน 50 คน

สมมติฐาน คือ ตัวแปรอิสระทั้ง 4
ตัวสามารถทำนายความน่ารักได้

การทดสอบสมมติฐาน

```
> ex2 <- read.table("lecture2ex2.csv", sep=",", header=TRUE)
> describe(ex2)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
face	1	50	51.64	15.26	51.50	51.02	15.57	21.00	89.00	68.00	0.27	-0.51	2.16
shape	2	50	50.96	16.75	49.00	50.75	15.57	13.00	99.00	86.00	0.25	0.09	2.37
habit	3	50	48.14	15.27	47.00	47.62	14.08	16.00	90.00	74.00	0.31	0.06	2.16
close	4	50	50.40	16.40	49.50	50.15	17.79	19.00	88.00	69.00	0.14	-0.67	2.32
cute	5	50	52.85	18.50	54.54	53.57	19.76	6.34	88.84	82.51	-0.37	-0.33	2.62

การทดสอบสมมติฐาน

```
> corr.test(ex2)
Call:corr.test(x = ex2)
Correlation matrix
       face shape habit close cute
face   1.00  0.45 -0.02  0.33 0.65
shape  0.45  1.00 -0.19  0.13 0.31
habit -0.02 -0.19  1.00  0.54 0.12
close  0.33  0.13  0.54  1.00 0.23
cute   0.65  0.31  0.12  0.23 1.00
Sample Size
[1] 50
Probability values (Entries above the diagonal are adjusted for multiple tests.)
      face shape habit close cute
face  0.00  0.01  1.00  0.13 0.00
shape 0.00  0.00  0.73  1.00 0.18
habit 0.90  0.18  0.00  0.00 1.00
close 0.02  0.37  0.00  0.00 0.54
cute  0.00  0.03  0.42  0.11 0.00

 To see confidence intervals of the correlations, print with the short=FALSE option
```

การทดสอบสมมติฐาน

```
> out2 <- lm(cute ~ face + shape + habit + close, data=ex2)
> summary(out2)

Call:
lm(formula = cute ~ face + shape + habit + close, data = ex2)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.372 -11.143   1.349   8.074  30.234 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.36760   10.90784   0.217    0.829    
face         0.80841    0.15839   5.104  6.5e-06 ***
shape        0.06468    0.13999   0.462    0.646    
habit        0.23651    0.16807   1.407    0.166    
close       -0.11788    0.16246  -0.726    0.472    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.27 on 45 degrees of freedom
Multiple R-squared:  0.4534,	Adjusted R-squared:  0.4048 
F-statistic: 9.332 on 4 and 45 DF,  p-value: 1.402e-05
```

$R^2 = .453$, $F(4, 45) = 9.33$, $p < .001$, Adjusted $R^2 = .405$, $s_e = 14.27$

การทดสอบสมมติฐาน

```
> out2 <- lm(cute ~ face + shape + habit + close, data=ex2)
> summary(out2)
```

$$\hat{Y} = 2.368 + 0.808X_1 + 0.065X_2 + 0.237X_3 - 0.118X_4$$

```
Residuals:
    Min      1Q  Median      3Q     Max 
-31.372 -11.143   1.349   8.074  30.234 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.36760   10.90784   0.217    0.829    
face         0.80841    0.15839   5.104  6.5e-06 ***
shape        0.06468    0.13999   0.462    0.646    
habit        0.23651    0.16807   1.407    0.166    
close       -0.11788    0.16246  -0.726    0.472    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.27 on 45 degrees of freedom
Multiple R-squared:  0.4534,    Adjusted R-squared:  0.4048 
F-statistic: 9.332 on 4 and 45 DF,  p-value: 1.402e-05
```

หน้าตามีผลต่อความน่ารักอย่างมีนัยสำคัญ ขณะที่ตัวแปรอื่นมีผลไม่ถึงระดับนัยสำคัญ

การทดสอบสมมติฐาน

```
> BetaDelta(out2)
Call:
BetaDelta(object = out2)

Standardized regression slopes with MVN standard errors:
          est     se       t df      p   0.05%    0.5%    2.5%  97.5%  99.5% 99.95%
face   0.6672 0.1036  6.4408 45 0.0000  0.3025  0.3886  0.4586 0.8758 0.9458 1.0319
shape  0.0586 0.1202  0.4872 45 0.6285 -0.3646 -0.2647 -0.1835 0.3007 0.3819 0.4817
habit  0.1952 0.1313  1.4868 45 0.1440 -0.2670 -0.1579 -0.0692 0.4597 0.5484 0.6575
close -0.1045 0.1365 -0.7655 45 0.4480 -0.5851 -0.4717 -0.3795 0.1705 0.2627 0.3761
> anova(out2)
Analysis of Variance Table

Response: cute
          Df Sum Sq Mean Sq F value    Pr(>F)    
face       1 7190.4  7190.4 35.3173 3.796e-07 ***
shape      1    3.6     3.6  0.0177    0.8947    
habit      1  298.8   298.8  1.4678    0.2320    
close      1  107.2   107.2  0.5264    0.4719    
Residuals 45 9161.8   203.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

การทดสอบสมมติฐาน

ท่านต้องการทดสอบว่าอะไร
ทำให้คนอื่นมองว่าคนน่ารัก
ท่านจึงทำนายความน่ารักด้วย
(1) หน้าตาดี (2) หุ่นดี (3) นิสัยดี
(4) ความใกล้ชิดสนิทสนม

สรุป

ตัวแปรอิสระทั้งหมดสามารถอธิบายการกระจายของความน่ารักได้ 45%

หน้าตาเป็นตัวบ่งชี้ความน่ารักเพียงอย่างเดียวที่ถึงระดับนัยสำคัญ

การเขียนรายงาน

- มี 2 แนวทางในการวิเคราะห์ถดถอย
  - เน้นการทำนายค่าของตัวแปรตาม การเขียนรายงานจะเน้นที่การสร้างสมการทำนาย และความแม่นยำของค่าที่ทำนายได้
  - เน้นการอธิบายอิทธิพลของตัวแปรอิสระ ค้นหาว่าตัวแปรใดที่กำหนดค่าของตัวแปรตาม การเขียนรายงานจะเน้นแค่ขนาดอิทธิพลและระดับนัยสำคัญของตัวแปรตาม
- งานวิจัยในจิตวิทยาจะมีทั้ง 2 รูปแบบ
  - ในงานวิจัยเชิงวิทยาศาสตร์บริสุทธิ์จะเน้นการอธิบายอิทธิพลของตัวแปรอิสระ (เช่น ตัวกำหนดความเครียด)
  - ในงานวิจัยเชิงประยุกต์บางกรณี (เช่น การคัดเลือกคนเข้าทำงาน การคัดกรองผู้ป่วย) จะเน้นการทำนายค่าตัวแปรตามโดยไม่สนใจอธิบายปรากฏการณ์

การเขียนรายงาน (เน้นทำนาย)

- ผู้วิจัยต้องการทำนายความเครียดจากความกดดันจากทางบ้านและความกดดันจากที่ทำงาน ตารางที่ 1 แสดงค่าสถิติพรรณนาและค่าสหสัมพันธ์ จากการวิเคราะห์ถดถอยแบบพหุ พบว่า ตัวแปรทั้งสองสามารถทำนายความเครียดได้อย่างมีนัยสำคัญ (R^2 = .36, p < .05) โดยสมการทำนาย คือ 16.011 + 0.459(บ้าน) + 0.255(ที่ทำงาน) โดยส่วนเบี่ยงเบนมาตรฐานของความผิดพลาดในการทำนายเท่ากับ 9.14 ซึ่งความชันของตัวแปรอิสระทั้งสองถึงระดับนัยสำคัญ (p < .05)

การเขียนรายงาน (เน้นทำนาย)

- ตารางที่ 1 แสดงสถิติพรรณนาและค่าสหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม (N = 100)

ตัวแปร	1	2	3
1. ความเครียด			
2. ความกดดันจากทางบ้าน	.568**		
3. ความกดดันจากที่ทำงาน	.481**	.554**	
M	52.10	50.25	51.03
SD	11.33	10.76	10.63

หมายเหตุ * $p < .05$, ** $p < .01$

การเขียนรายงาน (เน้นอธิบาย)

- งานวิจัยนี้ต้องการทดสอบว่าลักษณะใดของเพศหญิงที่มีผลทำให้ผู้อื่นมองว่าน่ารัก สถิติพรรณนาและค่าสหสัมพันธ์แสดงในตารางที่ 1 จากการสำรวจนิสิตระดับปริญญาตรีจำนวน 50 คน และวิเคราะห์ด้วยการวิเคราะห์ถดถอยแบบพหุ พบผลดังตารางที่ 2 ตัวแปรอิสระทั้งหมดสามารถทำนายความน่ารักได้อย่างมีนัยสำคัญ (R^2 = .45, p < .001) โดยหน้าตาเป็นตัวแปรเดียวที่สามารถทำนายความน่ารักได้อย่างมีนัยสำคัญ (p < .05)

การเขียนรายงาน (เน้นอธิบาย)

- ตารางที่ 1 แสดงสถิติพรรณนาและค่าสหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม (N = 50)

ตัวแปร	1	2	3	4	5
1. ความน่ารัก					
2. หน้าตา	.655**				
3. รูปร่าง	.306*	.447**			
4. นิสัย	.116	-.019	-.191		
5. ความใกล้ชิด	.230	.333**	.130	.537**	
M	52.85	51.64	50.96	48.14	50.40
SD	18.50	15.27	16.75	15.27	16.40

หมายเหตุ * p < .05, ** p < .01

การเขียนรายงาน (เน้นอธิบาย)

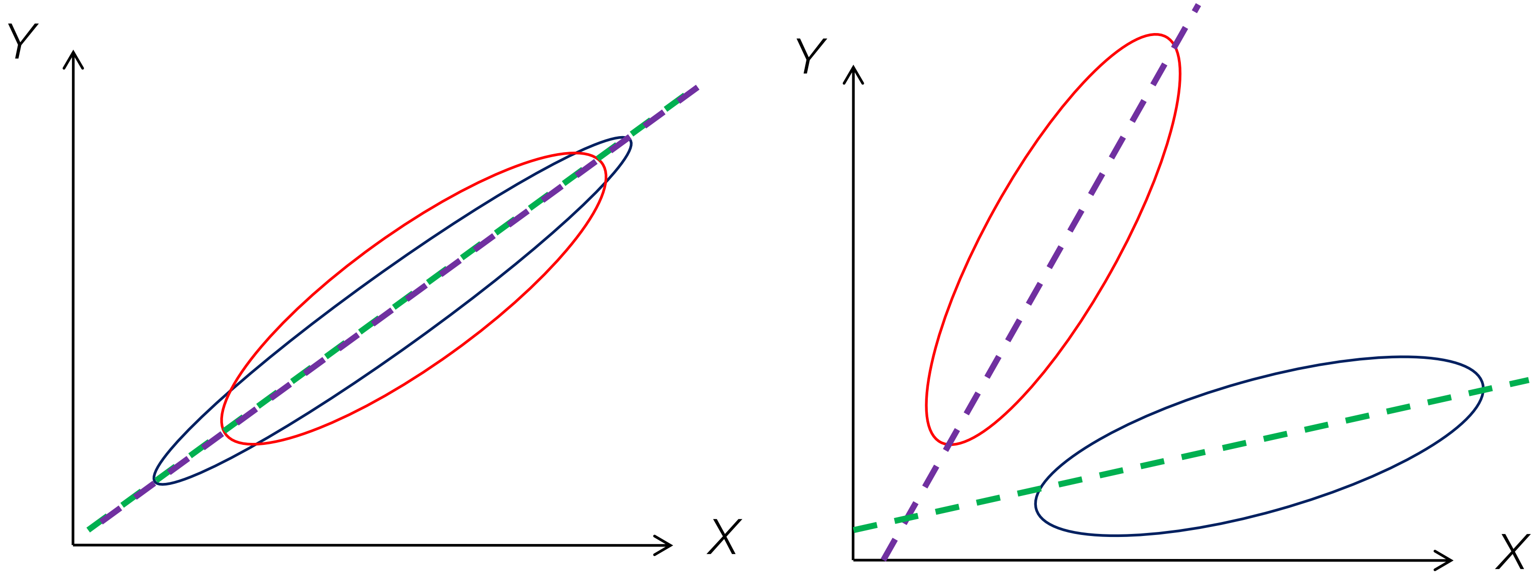
- ตารางที่ 2 แสดงผลการวิเคราะห์ถดถอย ในการทำนายความน่ารัก (N = 50)

ตัวแปร	b	SE	β
หน้าตา	0.808**	0.158	.667
รูปร่าง	0.065	0.140	.059
นิสัย	0.237	0.168	.195
ความใกล้ชิด	-0.118	0.162	-.105
R^2 = .45, Adjusted R^2 = .41, $p < .001$			

หมายเหตุ * $p < .05$, ** $p < .01$

ค่าสัมประสิทธิ์ถดถอยและค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสหสัมพันธ์ คือ การเกาะกลุ่ม
- ค่าสัมประสิทธิ์ถดถอย คือ อิทธิพลของตัวแปร X ที่มีต่อ Y



$r_1 \neq r_2; b_1 = b_2; a_1 = a_2$ $\quad$ $r_1 = r_2; b_1 \neq b_2; a_1 \neq a_2$

ค่าสัมประสิทธิ์ถดถอยและค่าสัมประสิทธิ์สหสัมพันธ์

- ค่าสหสัมพันธ์ เป็นสถิติที่สมมาตร
  - สหสัมพันธ์ระหว่าง X และ Y และสหสัมพันธ์ระหว่าง Y และ X เหมือนกัน
  - $r_{XY} = r_{YX}$
- ค่าสัมประสิทธิ์ถดถอย เป็นสถิติที่ไม่สมมาตร
  - ค่าความชันและจุดตัดแกน Y เมื่อทำนาย Y จาก X และทำนาย X จาก Y จะไม่เหมือนกัน
  - $b_{XY} \neq b_{YX}$; $a_{XY} \neq a_{YX}$
  - เนื่องจาก ทำนาย Y จาก X จะต้องแปลง X ไปใกล้ Y แต่เมื่อทำนาย X จาก Y จะต้องแปลง Y ไปใกล้ X

ค่าสัมประสิทธิ์ถดถอยและค่าสัมประสิทธิ์สหสัมพันธ์

- มีบางคนเข้าใจว่า การสหสัมพันธ์ไม่สามารถบอกความสัมพันธ์เชิงเหตุและผลได้ แต่การวิเคราะห์ถดถอยสามารถบอกความสัมพันธ์เชิงเหตุและผลได้
- ความเชื่อนี้ไม่ถูกต้องทั้งหมด เพราะความสัมพันธ์ก็สามารถใช้บอกความสัมพันธ์เชิงเหตุผลได้ (ดังที่กล่าวไปในเรื่องความสัมพันธ์)
    - ตัวแปร X มีความสัมพันธ์กับตัวแปร Y
    - X ต้องมาก่อน Y
    - การเกิด Y จะต้องไม่ได้รับการอธิบายด้วยตัวแปรอื่น
    - มีทฤษฎีหรือกระบวนการอธิบายว่าทำไม X ถึงก่อให้เกิด Y

ค่าสัมประสิทธิ์ถดถอยและค่าสัมประสิทธิ์สหสัมพันธ์

- แต่การวิเคราะห์ถดถอย เหมาะสมในการประกอบการอธิบายเชิงสาเหตุมากกว่า
  - การวิเคราะห์ถดถอยเป็นการมองให้ละเอียดมากขึ้น ว่าหากเหตุเพิ่มขึ้น 1 หน่วยแล้ว ผลเพิ่มขึ้นกี่หน่วย
  - นอกจากนี้ การวิเคราะห์ถดถอย สามารถบอกอิทธิพลของตัวแปรเหตุ เมื่อควบคุมตัวแปรควบคุมให้คงที่
- การวิเคราะห์ถดถอยก็เป็นเงื่อนไขที่จำเป็น แต่ไม่เพียงพอต่อการอธิบายเชิงสาเหตุ เช่น การเกิด Y จะต้องไม่ได้รับการอธิบายด้วยตัวแปรอื่น, X มาก่อน Y

ความสัมพันธ์เชิงเหตุผล

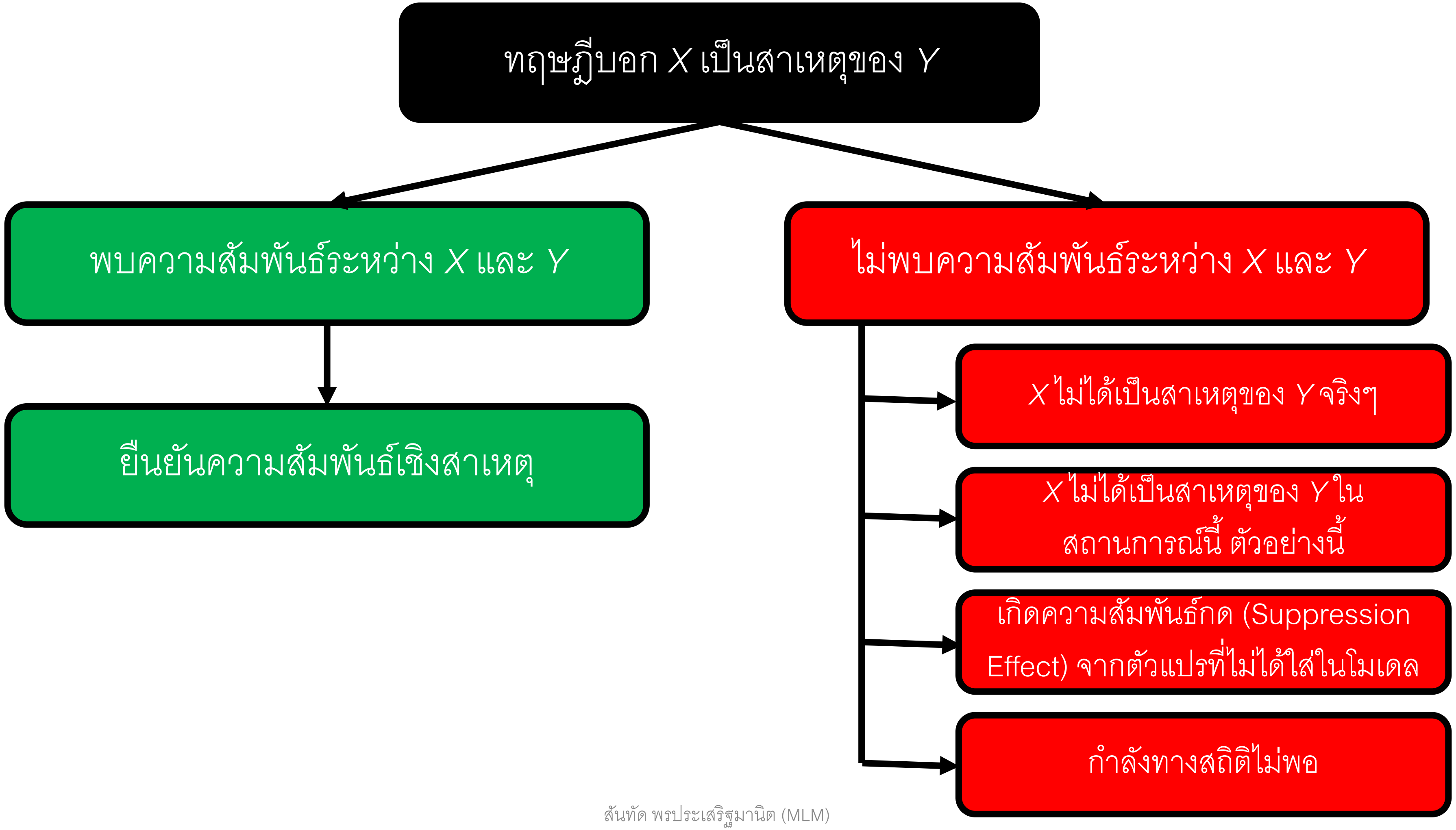
- การใช้คำว่า “อธิบาย” ก็เป็นการพูดถึงความสัมพันธ์เชิงเหตุและผล
- การวิเคราะห์ถดถอย, สหสัมพันธ์, หรือเปรียบเทียบค่าเฉลี่ย ใช้ในงานวิจัยที่มุ่งเน้นการอธิบายความสัมพันธ์เชิงเหตุผลได้ แต่การวิเคราะห์โดยตัวของมันเอง “*ไม่เพียงพอในการอ้างความสัมพันธ์เชิงเหตุผล*”

ความสัมพันธ์เชิงเหตุผล

- วิธีการในการหาความสัมพันธ์เชิงเหตุผล

องค์ประกอบ	ทฤษฎี	ตัวแปรควบคุม	จัดกระทำแบบสุ่ม
X มีความสัมพันธ์กับ *Y*	★★★	★★★	★★★
X มาก่อน *Y*	★	★	★★★
จำกัดอิทธิพลตัวแปรอื่น	-	★★	★★★
มีทฤษฎีรองรับ	★★★	★★★	★★

ความสัมพันธ์เชิงเหตุผล - ทฤษฎีอย่างเดียว

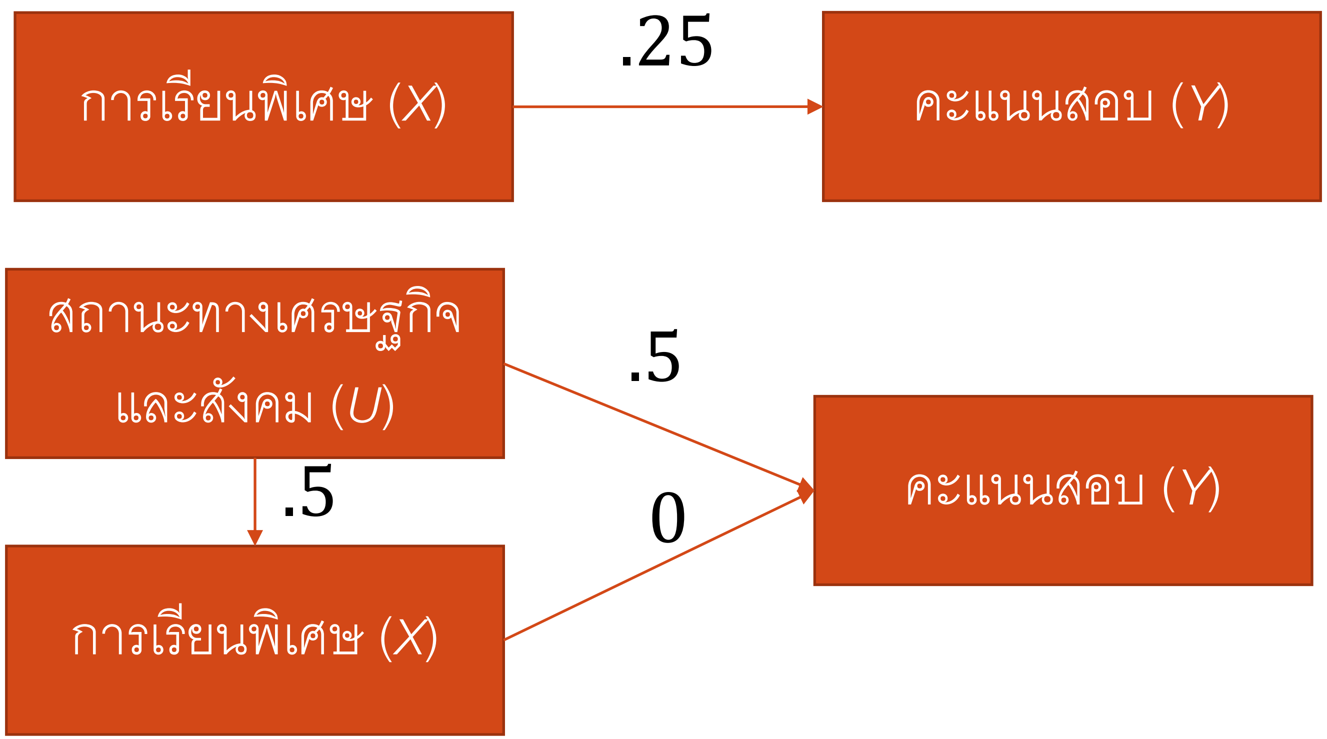


ความสัมพันธ์เชิงเหตุผล - ทฤษฎีอย่างเดียว

- การใช้ข้อมูลที่ไม่ใช้การทดลองมาหาความสัมพันธ์หรือความแตกต่าง แล้วอ้างความสัมพันธ์เชิงเหตุผล แม้จะทำได้ก็ตาม แต่เป็นวิธีการอ้างความสัมพันธ์เชิงเหตุผลที่อ่อนที่สุด
- กล่าวคือ เราอ้างเชิงสาเหตุจากทฤษฎีที่เราถือไว้ แต่เราจะแน่ใจได้อย่างไรว่า ทฤษฎีที่เรามีเป็นทฤษฎีที่เหมาะสมกับงานนี้ ไม่มีทฤษฎีอื่นมาหักล้างคำอธิบาย
  - เช่น ผมมีทฤษฎีเชื่อว่า “การกินยาฟ้าทะลายโจร ทำให้รักษาโควิดได้” ผมเลยเก็บข้อมูลเปรียบเทียบระหว่างคนที่กินและไม่กิน พบว่าคนกิน หายจากโควิดได้เร็วกว่า
  - ทฤษฎีอื่น ที่อาจมาอธิบายได้ เช่น คนที่กินจะมีแนวโน้มกินยาอื่นร่วมด้วย ทำให้ผลหายโควิดได้เร็วกว่า เกิดจากยาอื่น ไม่ใช่ผลจากยาฟ้าทะลายโจรโดยตรง
  - แต่จากวิธีอนุมานเหตุผลจากทฤษฎี ทำให้เรายืนยันทฤษฎีฟ้าทะลายโจรเลย

ความสัมพันธ์เชิงเหตุผล – ตัวแปรควบคุม

- สมมติว่ามีตัวแปร *U* ที่มีผลต่อทั้ง *X* และ *Y* จะส่งผลให้ความสัมพันธ์ระหว่าง *X* และ *Y* เปลี่ยนแปลงไป



ความสัมพันธ์เชิงเหตุผล – ตัวแปรควบคุม

- ตัวแปร *U* ในที่นี้จะเรียกว่าตัวแปรสาเหตุร่วม (Common causes) ซึ่งหากไม่ได้นำเข้ามาในการวิเคราะห์แล้ว ความสัมพันธ์ระหว่าง *X* และ *Y* จะไม่สะท้อนความเป็นจริง
- จากปรากฏการณ์นี้ สามารถสรุปได้ว่า ค่าสัมประสิทธิ์ถดถอย (หรือค่าความแตกต่างระหว่างกลุ่ม) จาก *X* ไปหา *Y* จะเปลี่ยนแปลงไป เมื่อชุดตัวแปรอิสระเปลี่ยนแปลงไป
- ด้วยเหตุนี้ ผู้วิเคราะห์จะต้องหาตัวแปรสาเหตุร่วมให้ได้ทั้งหมด กล่าวคือ
  - ต้อง**วัด**ตัวแปรสาเหตุร่วมทั้งหมด
  - ต้องนำตัวแปรสาเหตุร่วมมา**ใส่ในการวิเคราะห์**ข้อมูล เพื่อควบคุมทางสถิติ (Statistical Control) ซึ่งสามารถทำได้โดยการใช้การวิเคราะห์ถดถอยแบบพหุ หรือ ANCOVA

ความสัมพันธ์เชิงเหตุผล – ตัวแปรควบคุม

- การวัดตัวแปรสาเหตุร่วมทั้งหมดได้นั้น มีเงื่อนไข 3 ประการ
  - นักวิจัยต้องรู้ว่าตัวแปรสาเหตุร่วมมีอะไรบ้าง วิธีการที่ปลอดภัยที่สุด คือ นำตัวแปรสาเหตุร่วมมาจากทฤษฎีให้มากที่สุดเท่าที่ทำได้
  - นักวิจัยต้องสามารถวัดตัวแปรสาเหตุร่วมเหล่านั้นได้ ซึ่งบางตัวแปรอาจวัดไม่ได้ด้วยเวลาจำกัด หรือเป็นสิ่งที่วัดไม่ได้จริงๆ
  - ตัวแปรสาเหตุร่วมที่วัดได้นั้น ควรเป็นค่าที่เกิดขึ้นก่อนที่จะวัดตัวแปรอิสระและตัวแปรตาม เช่น เมื่อสามเดือนก่อนมีฐานะที่ดี (U_1) เลยสามารถให้ลูกเรียนพิเศษได้ (X_2) แต่ ณ ปัจจุบันฐานะไม่ได้ดีเช่นเดิม (U_2) แต่เนื่องจากลงทะเบียนเรียนพิเศษไปแล้ว เลยยังไปเรียน บางครั้งเราวัดได้เฉพาะ U_2 แต่วัด U_1 ไม่ได้
    - ถ้าตัวแปรสาเหตุร่วมเป็นตัวแปรที่คงที่พอสมควร อาจใช้ค่าปัจจุบันแทนได้

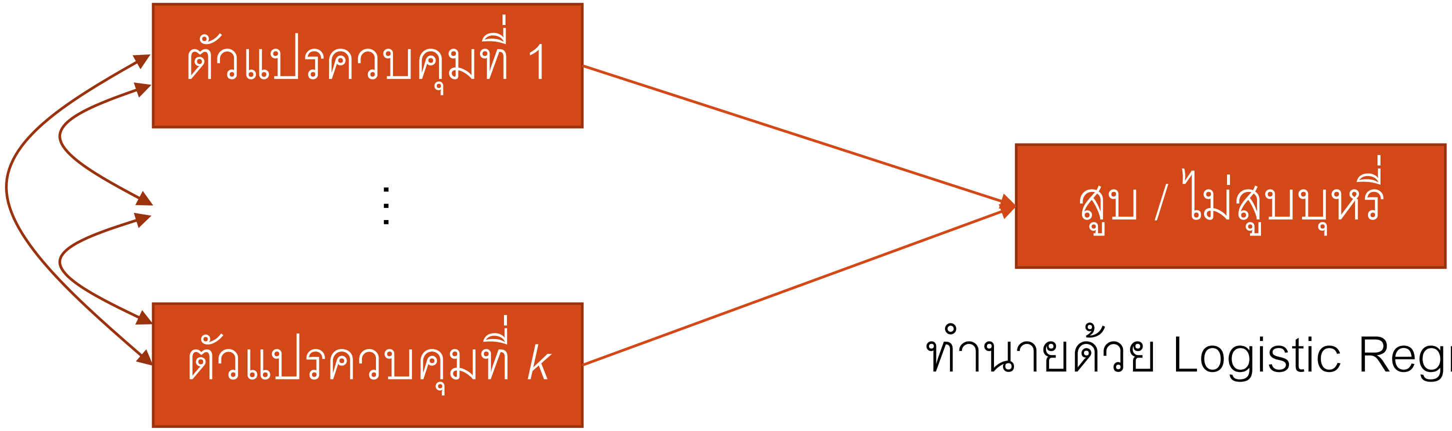
ความสัมพันธ์เชิงเหตุผล – การทดลองแบบสุ่ม

- วิธีการที่ง่ายที่สุด ในการตัดตัวแปรสาเหตุร่วมออกจากการวิเคราะห์ คือ การสุ่มคนให้อยู่ในค่าต่างๆ ของตัวแปรอิสระ
- เมื่อตัวแปรอิสระเกิดจากการสุ่มแล้ว ตัวแปรสาเหตุร่วมจะไม่มีอิทธิพลต่อตัวแปรอิสระเลย
- กล่าวคือ การสุ่มจะเป็นการขจัดอิทธิพลของตัวแปรแทรกซ้อนทั้งหมด ทั้งที่เรานึกถึง และที่เรานึกไม่ถึง
- ค่าสัมประสิทธิ์ถดถอย (หรือค่าความแตกต่าง) จาก X ไป Y จะสามารถแปลความหมายเป็นอิทธิพลเชิงเหตุผลได้โดยตรง
- ปัญหาของการทดลองแบบสุ่ม คือ เปลืองทรัพยากร และการสุ่มตัวแปรอิสระบางตัวอาจผิดหลักจริยธรรม (เช่น ผลของการสูบบุหรี่ต่อมะเร็ง)

ความสัมพันธ์เชิงเหตุผล – การจับคู่ตัวแปรควบคุม

- ถ้าเราสามารถจับคู่คนที่มีค่าตัวแปรควบคุมใกล้เคียงกันมากๆ ได้ แต่มีความแตกต่างในตัวแปรอิสระ (เช่น คู่หนึ่ง คนหนึ่งสูบบุหรี่ แต่คนหนึ่งไม่สูบบุหรี่ แต่มีตัวแปรอื่นเหมือนกันหมด ทั้งเพศ อายุ และอื่นๆ) ความแตกต่างของตัวแปรตามที่ได้ น่าจะผลของตัวแปรอิสระจริงๆ
- การจับคู่ (Matching) อาจทำได้ เมื่อตัวแปรควบคุมมีจำนวนน้อย แต่หากมีจำนวนมากแล้ว การจับคู่คนที่เหมือนกันแทบเป็นไปไม่ได้
- Rubin (1974) ได้เสนอวิธีคิดเรื่อง การจับคู่คะแนนโน้มเอียง (Propensity Score Matching) ขึ้น เพื่อยุบตัวแปรควบคุมจำนวนมากให้น้อยลงจนสามารถจับคู่ได้

ความสัมพันธ์เชิงเหตุผล – การจับคู่ตัวแปรควบคุม

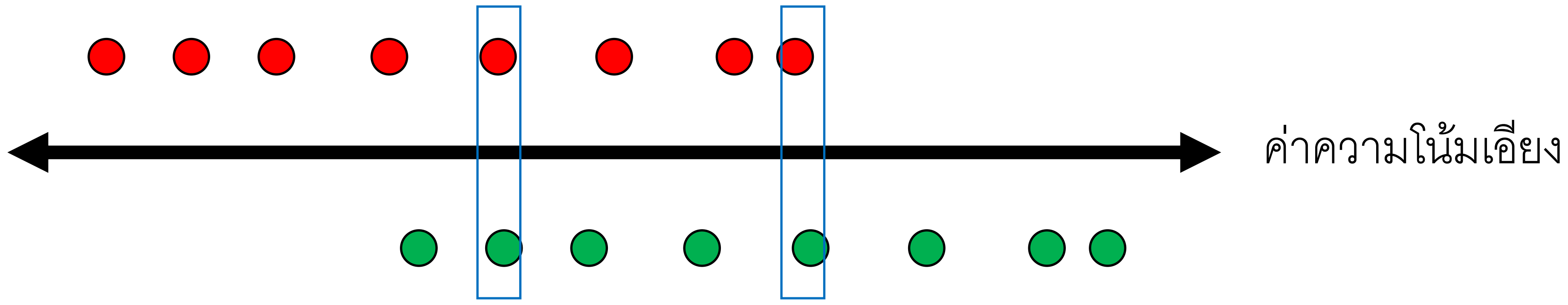


ทำนายด้วย Logistic Regression

ค่าที่ทำนายได้ (Predicted Score) จะเป็นค่าที่บอกว่า ถ้าพื้นฐานตัวแปรควบคุมมีค่าเท่านี้แล้วมีความโน้มเอียง (Propensity Score) ที่จะสูบบุหรี่เท่าไร

ความสัมพันธ์เชิงเหตุผล – การจับคู่ตัวแปรควบคุม

- จับคู่ของคนที่สูบบุหรี่และไม่สูบบุหรี่ที่มีค่าโน้มเอียงใกล้เคียงกัน ไม่เกินค่าจุดตัดค่าหนึ่ง คนที่จับคู่ไม่ได้จะคัดทิ้งออกจากการวิเคราะห์ต่อไป
- ความแตกต่างภายในคู่เดียวกันเฉลี่ยตลอดทุกคู่ จะเป็นผลลัพธ์ของตัวแปรอิสระที่มีต่อตัวแปรตาม



ความสัมพันธ์เชิงเหตุผล – การจับคู่ตัวแปรควบคุม

- วิธีการนี้เหมาะสมกับสถานการณ์ดังต่อไปนี้
  - มีตัวแปรควบคุมจำนวนมาก (100 ตัวแปรขึ้นไป)
  - ตัวแปรตาม เป็นเหตุการณ์ที่เกิดขึ้นยาก (Rare events) เช่น การเกิดมะเร็ง
  - การวิจัยมีการเก็บข้อมูลจำนวนมาก (เพราะกลุ่มตัวอย่างอาจถูกคัดทิ้งเยอะ)
- เงื่อนไขในหัวข้อการอนุมานเชิงเหตุผลด้วยตัวแปรควบคุมที่กล่าวไปข้างต้น ครอบคลุมถึงการใช้เทคนิคนี้
- ดูรายละเอียดเพิ่มเติมได้ที่ West & Thoemmes (2010)

การเปลี่ยนศูนย์กลาง

- การเปลี่ยนศูนย์กลาง (Centering) คือการนำตัวแปรอิสระไปลบด้วยค่าคงที่ เพื่อให้จุดตัดแกน Y มีความหมายในเชิงปฏิบัติ

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

ค่าที่ทำนายได้ เมื่อ X_1 และ X_2 เท่ากับ 0

$$\hat{Y} = a' + b_1(X_1 - c) + b_2(X_2 - d)$$

ค่าที่ทำนายได้ เมื่อ $X_1 = c$ และ $X_2 = d$

การเปลี่ยนศูนย์กลาง

- เช่น ทำนายผลการปฏิบัติงาน (M = 50, SD = 10) ด้วยคะแนนเชาวน์ปัญญา (M = 100, SD = 15) และคะแนนสอบ TOEIC (M = 550, SD = 100)

$$\boxed{\hat{Y} = 11.167 + 0.33(\mathrm{IQ}) + 0.01(\mathrm{TOEIC})}$$

ค่าผลการปฏิบัติงานที่ทำนายได้ เมื่อ IQ = 0 และ TOEIC = 0

$$\boxed{\hat{Y} = 49.667 + 0.33(\mathrm{IQ} - 100) + 0.01(\mathrm{TOEIC} - 550)}$$

ค่าผลการปฏิบัติงานที่ทำนายได้ เมื่อ IQ = 100 และ TOEIC = 550

จากตรงนี้ สามารถทดสอบว่าจุดตัดแกน Y แตกต่างจาก 0 อย่างมีนัยสำคัญหรือไม่

การเปลี่ยนศูนย์กลาง

- การเปลี่ยนศูนย์กลางจะสำคัญมากขึ้น หากผู้วิจัยต้องการใช้การวิเคราะห์ถดถอย ในการหาปฏิสัมพันธ์ระหว่างตัวแปร (Interaction) หรือการทำนายแบบไม่เป็นเส้นตรง (Nonlinear regression)
- การย้ายศูนย์กลางในการวิเคราะห์พหุระดับ เป็นสิ่งที่สำคัญมากต่อการแปลความหมายข้อมูล

การเปลี่ยนศูนย์กลาง

```
> out2a <- lm(cute ~ I(face - 50) + I(shape - mean(shape)) + I(habit - 50) + I(close - 50),
 data=ex2)
> summary(out2a)

Call:
lm(formula = cute ~ I(face - 50) + I(shape - mean(shape)) + I(habit -
    50) + I(close - 50), data = ex2)

Residuals:
    Min      1Q  Median      3Q     Max
-31.372 -11.143   1.349   8.074  30.234

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            52.01609    2.05732  25.283  < 2e-16 ***
I(face - 50)            0.80841    0.15839   5.104  6.5e-06 ***
I(shape - mean(shape))  0.06468    0.13999   0.462    0.646
I(habit - 50)           0.23651    0.16807   1.407    0.166
I(close - 50)          -0.11788    0.16246  -0.726    0.472
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.27 on 45 degrees of freedom
Multiple R-squared:  0.4534,    Adjusted R-squared:  0.4048
F-statistic: 9.332 on 4 and 45 DF,  p-value: 1.402e-05
```

สังเกตว่า ค่า t, F, R^2 ไม่เปลี่ยนแปลง

$$\hat{Y} = 52.02 + 0.808(X_1 - 50) + 0.065(X_2 - \bar{X}_2)$$
$$+0.237(X_3 - 50) - 0.118(X_4 - 50)$$

การเปลี่ยนศูนย์กลาง

สังเกตว่า ค่าต่างๆ ไม่เปลี่ยนแปลง

```
> BetaDelta(out2a)
Call:
BetaDelta(object = out2a)

Standardized regression slopes with MVN standard errors:
                         est     se       t df      p   0.05%    0.5%    2.5%  97.5%  99.5% 99.95%
I(face - 50)          0.6672 0.1036  6.4408 45 0.0000  0.3025  0.3886  0.4586 0.8758 0.9458 1.0319
I(shape - mean(shape)) 0.0586 0.1202  0.4872 45 0.6285 -0.3646 -0.2647 -0.1835 0.3007 0.3819 0.4817
I(habit - 50)         0.1952 0.1313  1.4868 45 0.1440 -0.2670 -0.1579 -0.0692 0.4597 0.5484 0.6575
I(close - 50)        -0.1045 0.1365 -0.7655 45 0.4480 -0.5851 -0.4717 -0.3795 0.1705 0.2627 0.3761
> anova(out2a)
Analysis of Variance Table

Response: cute
                       Df Sum Sq Mean Sq F value    Pr(>F)    
I(face - 50)            1 7190.4  7190.4 35.3173 3.796e-07 ***
I(shape - mean(shape))  1    3.6     3.6  0.0177    0.8947    
I(habit - 50)           1  298.8   298.8  1.4678    0.2320    
I(close - 50)           1  107.2   107.2  0.5264    0.4719    
Residuals              45 9161.8   203.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- ตัวแปรแบบนามบัญญัติ (Nominal scale) สามารถใช้เป็นตัวแปรอิสระได้ แต่ต้องแปลงให้เป็นตัวแปรแบบดัมมี่ (Dummy variable)
- ตัวแปรดัมมี่ คือ ตัวแปรที่มีค่า 1 หรือ 0 บ่งชี้ว่าเป็นสมาชิกกลุ่มที่กำหนดหรือไม่
- จำนวนของตัวแปรดัมมี่สำหรับตัวแปรแบบนามบัญญัติหนึ่ง จะเท่ากับจำนวนกลุ่มลบหนึ่ง

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- ตัวแปรแบบมี 2 กลุ่ม หรือที่เรียกว่าตัวแปรทวินาม (Dichotomous variable)

ตัวแปรนามบัญญัติ

การสัก
มีรอยสัก
ไม่มีรอยสัก

ตัวแปรดัมมี่

การสัก
1
0

- จำนวนตัวแปรดัมมี่ = 2 – 1 = 1 ตัวแปร
- กลุ่มอ้างอิง คือ กลุ่มที่ถูกกำหนดให้มีค่า 0 ในที่นี้คือ ไม่มีรอยสัก

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- ตัวแปรมีมากกว่า 2 กลุ่ม หรือที่เรียกว่าตัวแปรพหุนาม (Polytomous variable)

ตัวแปรนามบัญญัติ

ศาสนา
พุทธ
คริสต์
อิสลาม
อื่นๆ

ตัวแปรดัมมี่

พุทธ	คริสต์	อิสลาม
1	0	0
0	1	0
0	0	1
0	0	0

- จำนวนตัวแปรดัมมี่ = 4 – 1 = 3 ตัวแปร
- กลุ่มอ้างอิง คือ กลุ่มที่ถูกกำหนดให้มีค่า 0 ในที่นี้คือ กลุ่มศาสนาอื่นๆ

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- กลุ่มอ้างอิง สามารถเป็นกลุ่มใดก็ได้ตามที่ผู้วิจัยกำหนด โดยส่วนใหญ่แล้วจะใช้กลุ่มควบคุม เพื่อช่วยในการแปลความหมายสัมประสิทธิ์ถดถอย
  - จุดตัดแกน *Y* คือ ค่าเฉลี่ยของตัวแปรตามในกลุ่มอ้างอิง
  - ความชัน คือ ความแตกต่างระหว่างกลุ่มที่ใช้ในการกำหนดตัวแปรดัมมี่นั้น และกลุ่มอ้างอิง

ตัวแปรนามบัญญัติ	ตัวแปรดัมมี่
การสัก	**การสัก**
มีรอยสัก	1
ไม่มีรอยสัก	0

$$\hat{Y} = a + bX$$

ค่าที่ทำนายได้
ซึ่งค่าที่ดีที่สุด
คือค่าเฉลี่ยของกลุ่ม

ใส่ตัวแปรดัมมี่

$$\bar{Y}_{\text{สัก}} = a + b(1) = a + b$$
$$\bar{Y}_{\text{ไม่สัก}} = a + b(0) = a$$

$$a = \bar{Y}_{\text{ไม่สัก}}$$
$$b = \bar{Y}_{\text{สัก}} - \bar{Y}_{\text{ไม่สัก}}$$

ตัวแปรนามบัญญัติ

ศาสนา
พุทธ
คริสต์
อิสลาม
อื่นๆ

ตัวแปรดัมมี่

พุทธ	คริสต์	อิสลาม
1	0	0
0	1	0
0	0	1
0	0	0

$$\hat{Y} = a + \underbrace{b_1X_1 + b_2X_2 + b_3X_3}_{\text{ใส่ตัวแปรดัมมี่}}$$

$\hat{Y}$: ค่าเฉลี่ยของกลุ่ม

$$a = \bar{Y}_{\text{อื่นๆ}}$$
$$b_1 = \bar{Y}_{\text{พุทธ}} - \bar{Y}_{\text{อื่นๆ}}$$
$$b_2 = \bar{Y}_{\text{คริสต์}} - \bar{Y}_{\text{อื่นๆ}}$$
$$b_3 = \bar{Y}_{\text{อิสลาม}} - \bar{Y}_{\text{อื่นๆ}}$$

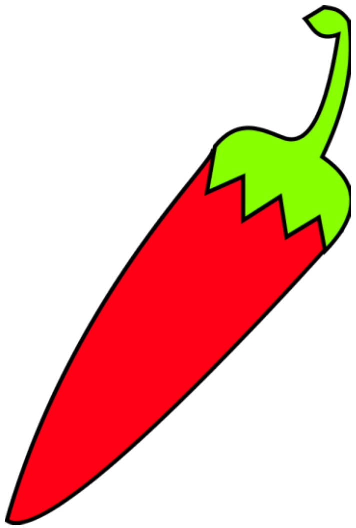
$$\bar{Y}_{\text{พุทธ}} = a + b_1(1) + b_2(0) + b_3(0) = a + b_1$$
$$\bar{Y}_{\text{คริสต์}} = a + b_1(0) + b_2(1) + b_3(0) = a + b_2$$
$$\bar{Y}_{\text{อิสลาม}} = a + b_1(0) + b_2(0) + b_3(1) = a + b_3$$
$$\bar{Y}_{\text{อื่นๆ}} = a + b_1(0) + b_2(0) + b_3(0) = a$$

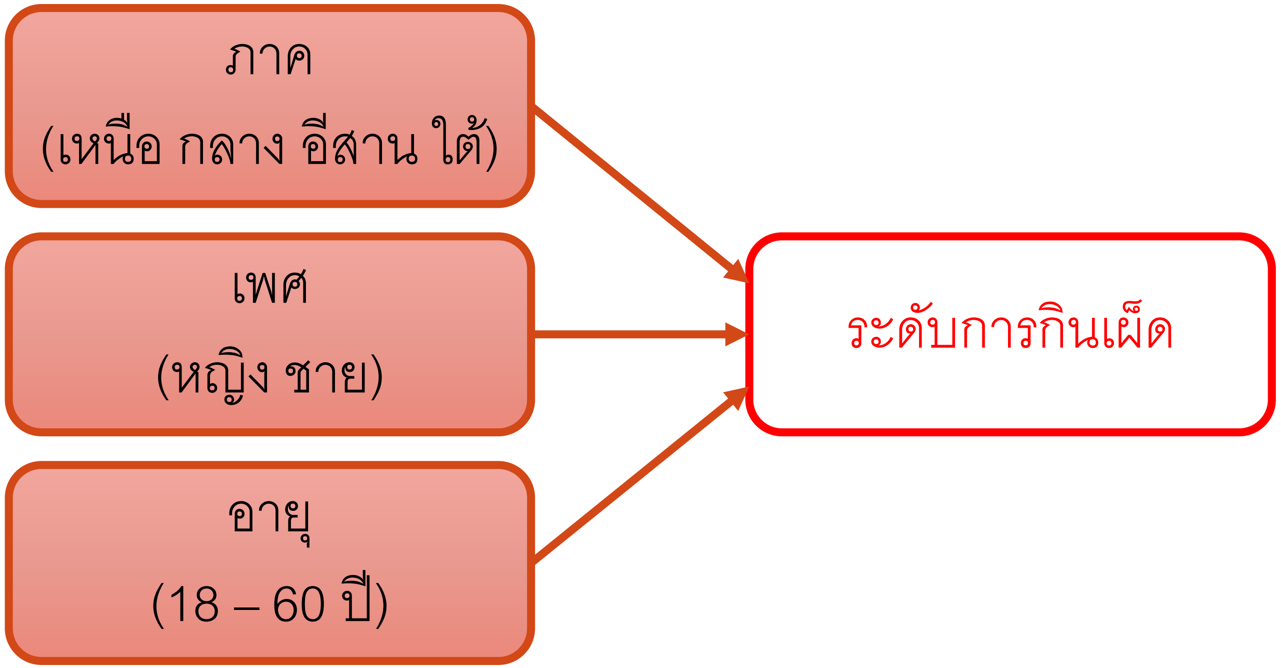
ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- แท้จริงแล้ว หากนำตัวแปรดัมมี่เป็นตัวแปรอิสระเพียงอย่างเดียว ผลที่ได้จะเทียบเท่ากับการวิเคราะห์ความแปรปรวนแบบทางเดียว
  - ค่า F จากการวิเคราะห์ถดถอยและความแปรปรวนทางเดียวจะเท่ากัน
  - หากมี 2 กลุ่ม ค่า t ในการทดสอบความชัน จะเท่ากับค่า t ในการทดสอบทีแบบอิสระจากกัน
  - ค่า R^2 เท่ากับค่า Eta-squared

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

หาปัจจัยที่อธิบายความแตกต่างระหว่างการกินเผ็ดโดยให้กลุ่มตัวอย่างปรุงรสก๋วยเตี๋ยวที่ตนเองทานเป็นประจำ แล้ววัดน้ำหนักของพริกป่นที่ใส่ โดยปัจจัยที่ทดสอบมีดังนี้





ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> ex3 <- read.table("lecture2ex3.csv", sep=",", header=TRUE)
> head(ex3, 15)
   id chilimg region sex age
1   1    4.08      1   1  23
2   2   11.15      3   2  43
3   3    7.75      2   2  36
4   4    0.00      1   1  39
5   5   20.09      4   2  24
6   6    6.78      2   2  30
7   7    2.36      2   1  59
8   8    6.18      2   1  45
9   9    1.81      2   2  43
10 10    0.00      1   1  52
11 11    7.47      3   1  21
12 12    0.05      2   2  56
13 13    1.24      2   1  18
14 14    0.00      2   2  50
15 15    6.89      1   2  20
```

ตัวอย่างข้อมูล

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> ex3[,"sex"] <- factor(ex3[,"sex"], labels=c("female", "male"))
> ex3[,"region"] <- factor(ex3[,"region"], labels=c("north", "center", "northeast", "south"))
> ex3[,"region"] <- relevel(ex3[,"region"], ref="south")
> head(ex3, 15)
   id chilimg    region    sex age
1   1    4.08     north female  23
2   2   11.15 northeast   male  43
3   3    7.75    center   male  36
4   4    0.00     north female  39
5   5   20.09     south   male  24
6   6    6.78    center   male  30
7   7    2.36    center female  59
8   8    6.18    center female  45
9   9    1.81    center   male  43
10 10    0.00     north female  52
11 11    7.47 northeast female  21
12 12    0.05    center   male  56
13 13    1.24    center female  18
14 14    0.00    center   male  50
15 15    6.89     north   male  20
```

แปลงค่าตัวเลขให้เป็นตัวแปรนามบัญญัติด้วยคำสั่ง `factor` โดยสามารถใส่ชื่อของแต่ละกลุ่มได้

R จะตั้งให้กลุ่มแรกเป็นกลุ่มอ้างอิงโดยอัตโนมัติ ถ้าต้องการเปลี่ยนกลุ่มอ้างอิงให้ใช้คำสั่ง `relevel`

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> describe(ex3)
        vars   n   mean     sd median trimmed    mad min    max  range  skew kurtosis   se
id         1 400 200.50 115.61 200.50  200.50 148.26   1 400.00 399.00  0.00    -1.21 5.78
chilimg    2 400   7.20   5.39   6.32    6.85   6.04   0  24.94  24.94  0.49    -0.68 0.27
region*    3 400   2.50   1.12   2.50    2.50   1.48   1   4.00   3.00  0.00    -1.37 0.06
sex*       4 400   1.50   0.50   1.50    1.50   0.74   1   2.00   1.00  0.00    -2.00 0.03
age        5 400  39.73  13.16  41.00   39.91  17.79  18  60.00  42.00 -0.10    -1.32 0.66
> with(ex3, table(sex))
sex
female   male
   200    200
> with(ex3, table(region))
region
    south     north    center northeast
      100       100       100       100
```

คำสั่ง describe จะปฏิบัติราวกับว่าตัวแปรทุกตัวเป็นตัวแปรต่อเนื่อง

ใช้คำสั่ง table เพื่อสร้างตารางแจกแจงความถี่ของตัวแปรนามบัญญัติ

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> corr.test(ex3[,c("chilimg", "age")])
Call:corr.test(x = ex3[, c("chilimg", "age")])
Correlation matrix
        chilimg   age
chilimg    1.00 -0.09
age       -0.09  1.00
Sample Size
[1] 400
Probability values (Entries above the diagonal are adjusted for multiple tests.)
        chilimg  age
chilimg    0.00 0.08
age        0.08 0.00

 To see confidence intervals of the correlations, print with the short=FALSE option
```

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~ I(age - 40) + sex + region, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)     -0.03047    0.01239  -2.459   0.0144 *  
sexmale          0.49914    0.32511   1.535   0.1255    
regionnorth    -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter    -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,    Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

$R^2 = .64, F(5, 394) = 140.2, p < .001$

ถ้าเป็นตัวแปรนามบัญญัติที่แปลงผ่านคำสั่ง `factor` แล้ว การวิเคราะห์ถดถอย จะสร้างตัวแปรดัมมี่โดยอัตโนมัติ

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)      -0.03047    0.01239  -2.459   0.0144 *  
sexmale           0.49914    0.32511   1.535   0.1255    
regionnorth     -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter     -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast  -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = 13.677 - 0.03(\text{Age} - 40) + 0.499(\text{Male})$$
$$-10.891(\text{North}) - 9.898(\text{Center}) - 6.139(\text{NE})$$

$$\hat{Y} = 13.677 - 0.03(Age - 40) + 0.499(Male) \\ -10.891(North) - 9.898(Center) - 6.139(NE)$$

เช่น หาค่าเฉลี่ยของแต่ละภาค เมื่ออายุเท่ากับ 40 ปี ในเพศชาย

$$\hat{Y} = 13.677 - 0.03(40 - 40) + 0.499(1) \\ -10.891(North) - 9.898(Center) - 6.139(NE)$$

$$\hat{Y} = 14.195 - 10.891(North) - 9.898(Center) - 6.139(NE)$$

$$\begin{aligned} \bar{Y}_{\text{North}} &= 14.195 - 10.891 = 3.304 \\ \bar{Y}_{\text{Center}} &= 14.195 - 9.898 = 4.297 \\ \bar{Y}_{\text{Northeast}} &= 14.195 - 6.139 = 8.056 \\ \bar{Y}_{\text{South}} &= 14.195 \end{aligned}$$

ตรงนี้เป็นค่าเฉลี่ยของแต่ละภาค
ในกลุ่มเพศชายอายุ 40 ปี

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)        -0.03047    0.01239  -2.459   0.0144 *  
sexmale             0.49914    0.32511   1.535   0.1255    
regionnorth       -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter       -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast    -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$a = 13.677$, $p < .001$	ค่าเฉลี่ยการใส่พริกป่นของหญิงใต้ที่อายุเท่ากับ 40 แตกต่างจาก 0 อย่างมีนัยสำคัญ
$b_{AGE} = -0.03$, $p = .014$	ผลของอายุเมื่อควบคุมเพศและภาค มีผลอย่างมีนัยสำคัญ
$b_{MALE} = 0.499$, $p = .126$	เพศชายและหญิงกินเผ็ดไม่แตกต่างกันอย่างมีนัยสำคัญ เมื่อควบคุมอายุและภาค
$b_{NORTH} = -10.891$, $p < .001$	ภาคเหนือกินเผ็ดน้อยกว่าภาคใต้อย่างมีนัยสำคัญ เมื่อควบคุมอายุและเพศ
$b_{CENTER} = -9.898$, $p < .001$	ภาคกลางกินเผ็ดน้อยกว่าภาคใต้อย่างมีนัยสำคัญ เมื่อควบคุมอายุและเพศ
$b_{NE} = -6.139$, $p < .001$	ภาคอีสานกินเผ็ดน้อยกว่าภาคใต้อย่างมีนัยสำคัญ เมื่อควบคุมอายุและเพศ

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

```
> anova(out3)
Analysis of Variance Table

Response: chilimg
              Df Sum Sq Mean Sq  F value    Pr(>F)    
I(age - 40)    1   86.2   86.22   8.1574  0.004516 ** 
sex            1   25.0   25.01   2.3661  0.124802    
region         3 7297.6 2432.52 230.1526 < 2.2e-16 ***
Residuals    394 4164.2   10.57                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ดังที่กล่าวไปข้างต้น คือ อย่าใช้ตารางนี้ ตารางนี้เทียบได้กับ Type I SS ใน ANOVA ซึ่งการแปลความหมายยาก ให้ใช้การวิเคราะห์ถดถอยแบบลำดับขั้น (Hierarchical Regression) ในการวิเคราะห์ผลของตัวแปรนามบัญญัติทั้งตัวแปร เช่น ผลของภาคที่มีต่อการกินเผ็ด

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- ค่าเฉลี่ยที่คำนวณได้จากสมการถดถอย เมื่อแทนค่าตัวแปรอื่นเป็นค่าใดค่าหนึ่ง อาจเรียกว่า ค่าเฉลี่ยที่ปรับแล้ว (Adjusted means)
- แตกต่างกับ ค่าเฉลี่ยที่หาจากคะแนนของแต่ละกลุ่มโดยตรง ที่เรียกว่า ค่าเฉลี่ยที่ยังไม่ได้ปรับ (Unadjusted means)

ตัวแปรจัดกลุ่มเป็นตัวแปรอิสระ

- การจัดการตัวแปรจัดอันดับ (Ordinal scale) นั้น สามารถทำได้ 2 รูปแบบ
  - ใส่ตัวเลข แทนแต่ละอันดับ แล้วนำตัวเลขใช้เป็นตัวแปรอิสระเลย

ความพึงพอใจในสินค้า	น้อยมาก	น้อย	ปานกลาง	สูง	สูงมาก
ตัวเลขที่ใช้แทน	1	2	3	4	5

ตัวเลขที่ใช้ ขึ้นอยู่กับวิจารณญาณของผู้วิจัย

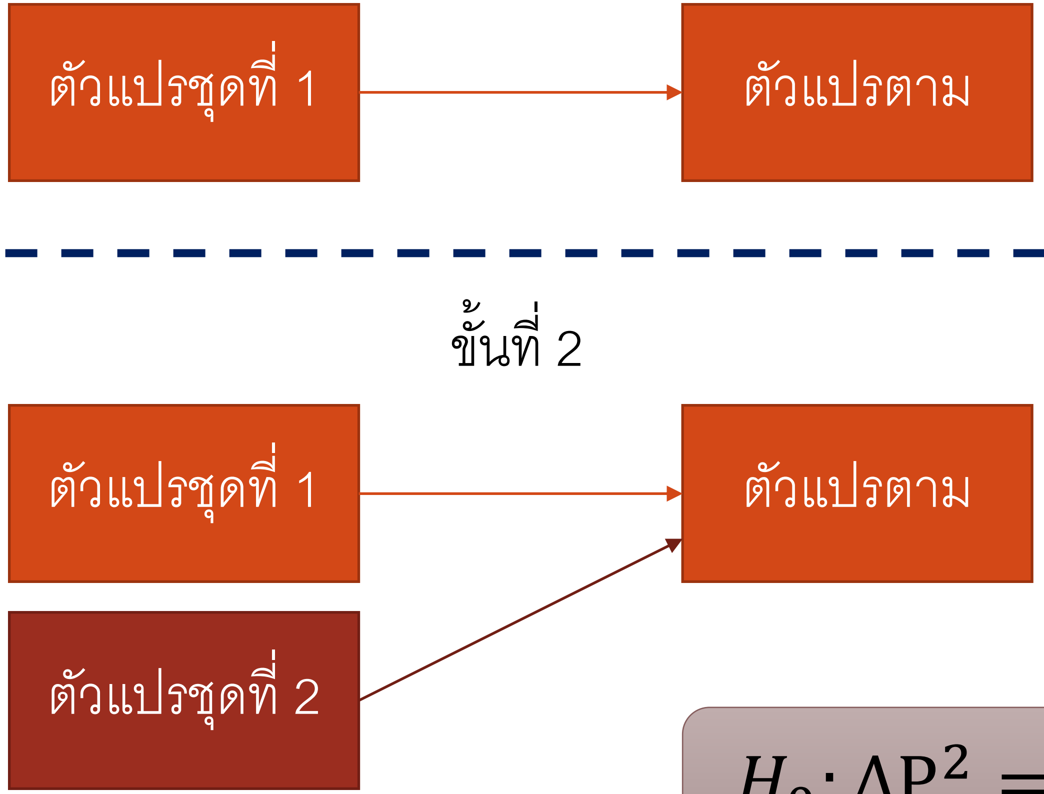
  - ไม่สนใจอันดับ แปลงเป็นตัวแปรดัมมี่เลย
- ถ้าเป็นไปได้ ให้ใช้วิธีแรก แต่หากไม่แน่ใจว่าการใส่ตัวเลขแทนถูกต้องหรือไม่ ให้ใช้การแปลงตัวแปรดัมมี่

การวิเคราะห์ถดถอยแบบลำดับขั้น

- ตัวแปรอิสระไม่จำเป็นต้องใส่พร้อมกันทีเดียว อาจจะใส่ทีละชุดได้
- เช่น นักวิจัยต้องการทดสอบอิทธิพลของบุคลิกภาพ ต่อผลการปฏิบัติงาน แต่ทราบว่าอายุงานและระดับการศึกษามีผลต่อผลการปฏิบัติงานด้วย
  - ขั้นที่ 1 ใส่อายุงานและระดับการศึกษาเป็นตัวแปรอิสระ
  - ขั้นที่ 2 ใส่บุคลิกภาพเป็นตัวแปรอิสระเพิ่มเติม
- วิธีการแบบนี้ ทำให้ผู้วิจัยสามารถเปรียบเทียบสัมประสิทธิ์การทำนายที่เพิ่มขึ้นได้
- วิธีการแบบนี้ จะเรียกว่า การวิเคราะห์ถดถอยแบบลำดับขั้น (Hierarchical Regression, Sequential Regression)

การวิเคราะห์ถดถอยแบบลำดับขั้น

ขั้นที่ 1



ขั้นที่ 2

R_A^2

R_{A+B}^2

ΔR^2

ใช้ *F* test ทดสอบว่า
ตัวแปรที่ใส่เข้าไปมีผล
ทำนายตัวแปรตาม
เพิ่มขึ้นหรือไม่

$H_0\colon \Delta \mathrm{P}^2 = 0$

$H_1\colon \Delta \mathrm{P}^2 > 0$

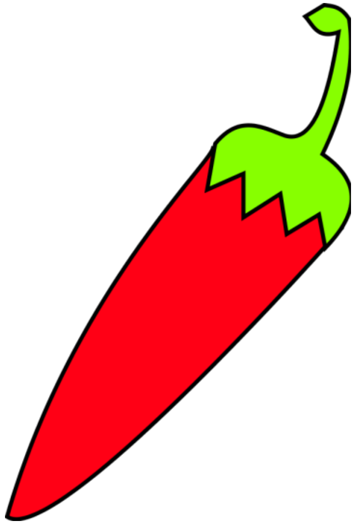
การวิเคราะห์ถดถอยแบบลำดับขั้น

ประโยชน์ของการวิเคราะห์ถดถอยแบบลำดับขั้น

- ผู้วิจัยสามารถใส่ตัวแปรควบคุม หรือตัวแปรที่มีงานวิจัยสนับสนุนจำนวนมากแล้วในขั้นที่ 1 หลังจากนั้นให้ใส่ตัวแปรที่ผู้วิจัยสนใจทดสอบในขั้นที่ 2
  - เช่น เชาวน์ปัญญาเป็นตัวแปรอิสระที่ใช้ในการคัดเลือกบุคลากรเพื่อเข้าทำงาน ผู้วิจัยสร้างวิธีการสัมภาษณ์ใหม่ขึ้นมา เพื่อดูว่าสามารถทำนายได้มากขึ้นกว่าเดิมหรือไม่
- หากผู้วิจัย มีตัวแปรอิสระที่สามารถจับกลุ่มเป็นชุดเดียวกัน แล้วต้องการทดสอบว่าตัวแปรอิสระชุดนี้มีอิทธิพลต่อตัวแปรตามหรือไม่ หากควบคุมตัวแปรอิสระอื่น
  - เช่น ตัวแปรบุคลิกภาพ 5 องค์ประกอบ ตัวแปรดัมมี่จากตัวแปรนามบัญญัติเดียวกัน

การวิเคราะห์ถดถอยแบบลำดับขั้น

ต้องการทดสอบว่าภาคถิ่นกำเนิด สามารถทำนายการทานอาหารเผ็ดเพิ่มขึ้นจากเพศและอายุหรือไม่



สมมติฐาน คือ ภาคสามารถอธิบายความแปรปรวนได้สูงขึ้น

H_0: $\Delta P^2 = 0$

H_1: $\Delta P^2 > 0$

ขั้นที่ 1: เพศ, อายุ

ขั้นที่ 2: เพศ, อายุ, และภาค

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> out3noregion <- lm(chilimg ~ I(age - 40) + sex, data=ex3)
> summary(out3noregion)

Call:
lm(formula = chilimg ~ I(age - 40) + sex, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2233 -4.3438 -0.8815  3.7121 16.9294 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.94329    0.38001  18.271   <2e-16 ***
I(age - 40) -0.03545    0.02044  -1.735   0.0836 .  
sexmale      0.50009    0.53733   0.931   0.3526    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.373 on 397 degrees of freedom
Multiple R-squared:  0.009611,	Adjusted R-squared:  0.004621 
F-statistic: 1.926 on 2 and 397 DF,  p-value: 0.1471
```

```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~ I(age - 40) + sex + region, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)     -0.03047    0.01239  -2.459   0.0144 *  
sexmale          0.49914    0.32511   1.535   0.1255    
regionnorth    -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter    -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

$R^2 = .01$, $F(2, 397) = 1.93$, $p = .15$

$R^2 = .64$, $F(5, 394) = 140.2$, $p < .001$

$\Delta R^2 = .63$

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> anova(out3noregion, out3)
Analysis of Variance Table

Model 1: chilimg ~ I(age - 40) + sex
Model 2: chilimg ~ I(age - 40) + sex + region
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    397 11461.8
2    394  4164.2  3    7297.6 230.15 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ΔR^2 = .63, $F(3, 394) = 230.15$, $p < .001$

ภาคสามารถอธิบายการกินเผ็ดได้เพิ่มเติมจากอายุและเพศอย่างมีนัยสำคัญ

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> out3nosex <- lm(chilimg ~ I(age - 40) + region, data=ex3)
> anova(out3nosex, out3)
Analysis of Variance Table

Model 1: chilimg ~ I(age - 40) + region
Model 2: chilimg ~ I(age - 40) + sex + region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    395 4189.2
2    394 4164.2  1    24.913 2.3571 0.1255
> out3noage <- lm(chilimg ~ sex + region, data=ex3)
> anova(out3noage, out3)
Analysis of Variance Table

Model 1: chilimg ~ sex + region
Model 2: chilimg ~ I(age - 40) + sex + region
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    395 4228.2
2    394 4164.2  1    63.907 6.0465 0.01436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F(1, 394) = 2.36$, $p = .13$

อายุไม่สามารถอธิบายการกินเผ็ดได้เพิ่มเติมจากภาคและเพศได้อย่างมีนัยสำคัญ

$F(1, 394) = 6.05$, $p = .014$

เพศสามารถอธิบายการกินเผ็ดได้เพิ่มเติมจากอายุและภาคอย่างมีนัยสำคัญ

เมื่อ F มี $df = 1$ แล้ว ค่า t^2 จากโมเดลที่ใส่ตัวแปรทุกตัวจะเท่ากับ F และจะพบว่าค่า p ที่ได้มีค่าเท่ากัน

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> BetaDelta(out3)
Call:
BetaDelta(object = out3)

Standardized regression slopes with MVN standard errors:
                    est     se        t  df      p   0.05%    0.5%    2.5%   97.5%   99.5%  99.95%
I(age - 40)     -0.0745 0.0301  -2.4725 394 0.0138 -0.1744 -0.1525 -0.1337 -0.0153  0.0035  0.0254
sexmale          0.0464 0.0300   1.5457 394 0.1230 -0.0531 -0.0313 -0.0126  0.1054  0.1241  0.1459
regionnorth     -0.8768 0.0389 -22.5656 394 0.0000 -1.0056 -0.9773 -0.9532 -0.8004 -0.7762 -0.7480
regioncenter    -0.7968 0.0401 -19.8661 394 0.0000 -0.9298 -0.9006 -0.8756 -0.7179 -0.6930 -0.6638
regionnortheast -0.4942 0.0390 -12.6839 394 0.0000 -0.6234 -0.5950 -0.5708 -0.4176 -0.3933 -0.3650
```

สำหรับตัวแปรแบบจัดกลุ่ม จะแปลความหมายสัมประสิทธิ์ถดถอยมาตรฐานค่อนข้างยาก (แม้ว่าในตัวอย่างนี้ จะดูง่ายว่าภาคมีผลสูงสุด) โดยทั่วไป ควรใช้การเปลี่ยนแปลงของ R^2 เปรียบเทียบระหว่างเมื่อตัวแปรอยู่และไม่อยู่ในโมเดล ตัวแปรใดเกิดการเปลี่ยนแปลงของ R^2 มากที่สุด แสดงว่ามีอิทธิพลต่อตัวแปรตามสูงที่สุด

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> summary(out3)$r.squared - summary(out3noage)$r.squared
[1] 0.005522048
> summary(out3)$r.squared - summary(out3nosex)$r.squared
[1] 0.002152659
> summary(out3)$r.squared - summary(out3noregion)$r.squared
[1] 0.6305658
```

ภาคทำให้ค่า R^2 เพิ่มถึง 63% ขณะที่เพศและอายุเพิ่ม R^2 ไม่ถึง 1%

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> out3noregion <- lm(chilimg ~ I(age - 40) + sex, data=ex3)
> summary(out3noregion)

Call:
lm(formula = chilimg ~ I(age - 40) + sex, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2233 -4.3438 -0.8815  3.7121 16.9294 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.94329    0.38001  18.271   <2e-16 ***
I(age - 40) -0.03545    0.02044  -1.735   0.0836 .  
sexmale      0.50009    0.53733   0.931   0.3526    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.373 on 397 degrees of freedom
Multiple R-squared:  0.009611,	Adjusted R-squared:  0.004621 
F-statistic: 1.926 on 2 and 397 DF,  p-value: 0.1471
```

```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~ I(age - 40) + sex + region, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)      -0.03047    0.01239  -2.459   0.0144 *  
sexmale           0.49914    0.32511   1.535   0.1255    
regionnorth     -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter     -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast  -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

$$\text{Model 1: } \hat{Y} = 6.943 - 0.035(\text{Age} - 40) + 0.500(\text{Male})$$

$$\text{Model 2: } \hat{Y} = 13.677 - 0.030(\text{Age} - 40) + 0.499(\text{Male}) - 10.891(\text{North}) - 9.898(\text{Center}) - 6.139(\text{NE})$$

การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> out3noregion <- lm(chilimg ~ I(age - 40) + sex, data=ex3)
> summary(out3noregion)

Call:
lm(formula = chilimg ~ I(age - 40) + sex, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2233 -4.3438 -0.8815  3.7121 16.9294 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.94329    0.38001  18.271   <2e-16 ***
I(age - 40) -0.03545    0.02044  -1.735   0.0836 .  
sexmale      0.50009    0.53733   0.931   0.3526    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.373 on 397 degrees of freedom
Multiple R-squared:  0.009611,	Adjusted R-squared:  0.004621 
F-statistic: 1.926 on 2 and 397 DF,  p-value: 0.1471
```

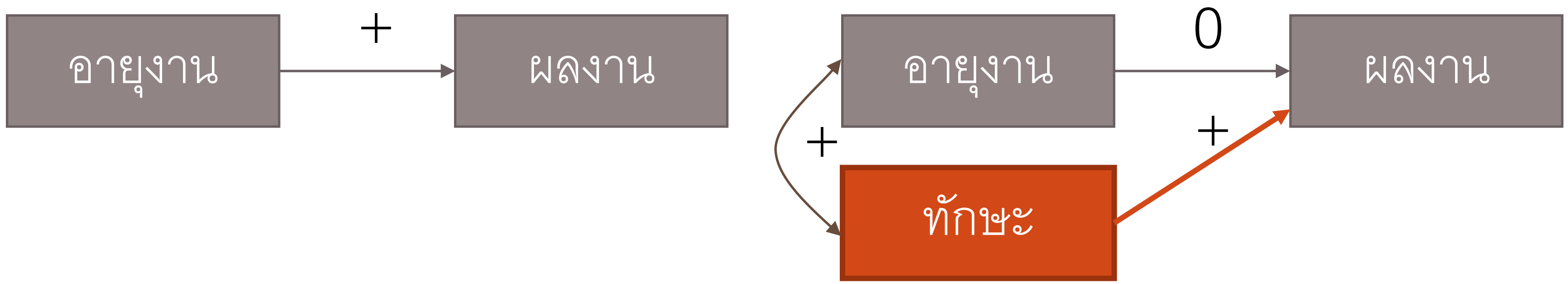
```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~ I(age - 40) + sex + region, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)      -0.03047    0.01239  -2.459   0.0144 *  
sexmale           0.49914    0.32511   1.535   0.1255    
regionnorth     -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter     -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast  -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

ค่าความชันมากขึ้น → เกิดผลของการกด (Suppression Effect)

ค่าความชันลดลง → อิทธิพลของตัวแปรลดลง เป็นเหตุการณ์ที่เกิดตามปกติ

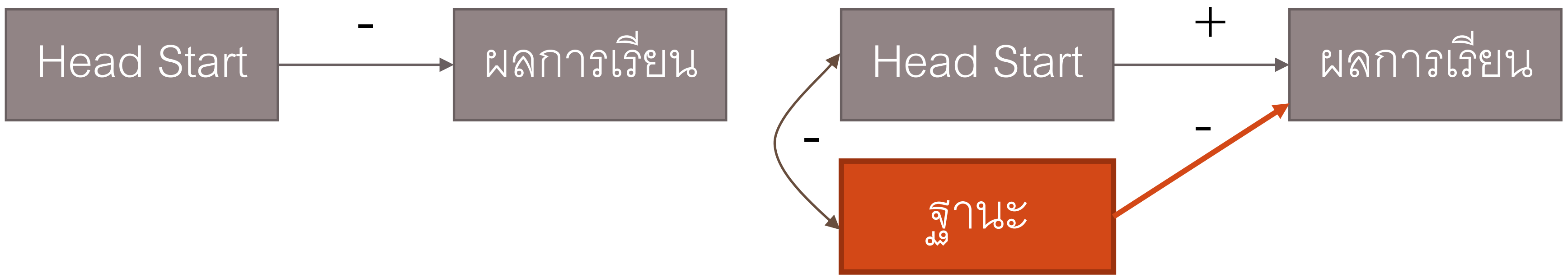
การวิเคราะห์ถดถอยแบบลำดับขั้น

- ค่าของความชันลดลง เมื่อมีตัวแปรอื่นร่วมในการถดถอย เป็นเหตุการณ์ที่เกิดขึ้นเป็นปกติในการวิเคราะห์ถดถอย
  - เช่น ท่านพบว่าอายุงานมีผลต่อผลการปฏิบัติงาน
  - แต่คนที่อายุงานมาก มักจะมีทักษะในการทำงานมากกว่า
  - ดังนั้น หากทำนายผลการปฏิบัติงาน ด้วยอายุงานและทักษะพร้อมกัน ค่าสัมประสิทธิ์ถดถอยของอายุงานน้อยลง (จนบางครั้ง ไม่ถึงระดับนัยสำคัญ)



การวิเคราะห์ถดถอยแบบลำดับขั้น

- ผลของการกด (Suppression Effect) คือ ผลของตัวแปรเด่นชัดขึ้น ถ้ามีตัวแปรอื่นร่วมในการถดถอย
  - เช่น คนที่เข้าร่วมโปรแกรม Head Start มีแนวโน้มที่จะมีผลการเรียนน้อยกว่าคนที่ไม่เข้าร่วม (ดูแต่เพียงค่าสหสัมพันธ์ระหว่าง 2 ตัวแปร)
  - แต่คนที่เข้าร่วมโปรแกรม Head Start ก็มักจะมีสถานะทางเศรษฐกิจไม่ดี
  - ดังนั้น ผลของโปรแกรม Head Start มีผลทางบวกต่อผลการเรียน หากควบคุมสถานทางเศรษฐกิจ



การวิเคราะห์ถดถอยแบบลำดับขั้น

```
> out3noregion <- lm(chilimg ~ I(age - 40) + sex, data=ex3)
> summary(out3noregion)

Call:
lm(formula = chilimg ~ I(age - 40) + sex, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2233 -4.3438 -0.8815  3.7121 16.9294 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.94329    0.38001  18.271   <2e-16 ***
I(age - 40) -0.03545    0.02044  -1.735   0.0836 .  
sexmale      0.50009    0.53733   0.931   0.3526    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.373 on 397 degrees of freedom
Multiple R-squared:  0.009611,	Adjusted R-squared:  0.004621 
F-statistic: 1.926 on 2 and 397 DF,  p-value: 0.1471
```

```
> out3 <- lm(chilimg ~ I(age - 40) + sex + region, data=ex3)
> summary(out3)

Call:
lm(formula = chilimg ~ I(age - 40) + sex + region, data = ex3)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2202 -2.5846 -0.0098  2.0427 10.2763 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.67700    0.36355  37.621   <2e-16 ***
I(age - 40)      -0.03047    0.01239  -2.459   0.0144 *  
sexmale           0.49914    0.32511   1.535   0.1255    
regionnorth     -10.89126    0.45981 -23.686   <2e-16 ***
regioncenter     -9.89767    0.46027 -21.504   <2e-16 ***
regionnortheast  -6.13874    0.45978 -13.351   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.251 on 394 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.6356 
F-statistic: 140.2 on 5 and 394 DF,  p-value: < 2.2e-16
```

ในที่นี้ ความชันไม่ได้เปลี่ยนแปลง จนอธิบายผลของอายุและเพศเพิ่มเติมได้

การวิเคราะห์ถดถอยแบบลำดับขั้น

- แท้จริงแล้ว หากตัวแปรที่ใส่เพิ่มในขั้นที่ 2 มีเพียงแค่ตัวแปรเดียว ค่า t ในการทดสอบความชันของตัวแปรนั้น หากนำมายกกำลังสองแล้วจะเท่ากับค่า F เลย และค่า p ที่ได้จะเท่ากันด้วย
- ด้วยเหตุนี้ การวิเคราะห์ถดถอยแบบลำดับขั้น จะใช้ก็ต่อเมื่อ ตัวแปรอิสระที่เพิ่มมีมากกว่า 1 ตัว
- ในบางกรณี ผู้วิจัยต้องการใส่ตัวแปรอิสระมากกว่า 2 ลำดับขั้น เช่น
    - ขั้นที่ 1 ใส่ตัวแปรเพศลงไป เพื่อทำนายความเผ็ด
    - ขั้นที่ 2 ใส่ตัวแปรอายุลงไปทำนายเพิ่มเติม
    - ขั้นที่ 3 ใส่ตัวแปรแสดงปฏิสัมพันธ์ระหว่างเพศและอายุลงไป

การวิเคราะห์ถดถอยแบบลำดับขั้น (การเขียนรายงาน)

- งานวิจัยนี้ต้องการทดสอบว่าเพศ (ชาย 50%, หญิง 50%), อายุ (*M* = 39.73 ปี, *SD* = 13.16 ปี), และภาค (เหนือ 25%, กลาง 25%, อีสาน 25%, ใต้ 25%) เป็นปัจจัยในการทำนายการกินเผ็ดของคนไทย (*M* = 7.2 กรัม, *SD* = 5.39 กรัม) หรือไม่ จึงเก็บกลุ่มตัวอย่างจำนวน 400 คน และวิเคราะห์ด้วยการวิเคราะห์ถดถอยแบบพหุ โดยแปลงเพศและภาค เป็นตัวแปรแบบดัมมี่ ตารางที่ 1 แสดงค่าสหสัมพันธ์ระหว่างตัวแปรในการวิเคราะห์ถดถอย ผลการทดสอบพบว่าตัวแปรทั้งหมดสามารถทำนายการกินเผ็ดได้อย่างมีนัยสำคัญ, $R^2 = .64$, $F(5, 364) = 140.2$, $p < .001$

การวิเคราะห์ถดถอยแบบลำดับขั้น (การเขียนรายงาน)

- ตารางที่ 1 แสดงค่าสหสัมพันธ์ระหว่างตัวแปรในการวิเคราะห์ถดถอย ($N = 400$)

ตัวแปร	1	2
1. การกินเผ็ด		
2. อายุ	-.086*	
M	7.20	39.74
SD	5.39	13.16

หมายเหตุ *$p < .05$, **$p < .01$

การวิเคราะห์ถดถอยแบบลำดับขั้น (การเขียนรายงาน)

- โดยอายุเป็นปัจจัยทางลบอย่างมีนัยสำคัญ, b = -0.03, p = .014, ปัจจัยเพศไม่ถึงระดับนัยสำคัญ, b = 0.50, p = .126, ส่วนภาค เนื่องจากว่าต้องทดสอบตัวแปรดัมมี่ 3 ตัวแปรพร้อมกัน จึงทดสอบด้วยการวิเคราะห์ถดถอยแบบลำดับขั้น พบว่าภาคเป็นปัจจัยทำนายการกินเผ็ดอย่างมีนัยสำคัญ, $\Delta R^2 = .63$, $F(3, 394) = 230.15$, $p < .001$, โดยภาคเหนือ, ภาคกลาง, และภาคอีสาน กินเผ็ดน้อยกว่าภาคใต้อย่างมีนัยสำคัญ ($p < .001$) ผลการวิเคราะห์ถดถอยแสดงไว้ดังตารางที่ 2

การวิเคราะห์ถดถอยแบบลำดับขั้น (การเขียนรายงาน)

- ตารางที่ 2 ผลการวิเคราะห์ถดถอยแบบลำดับขั้น ในการทำนายการกินเผ็ดของกลุ่มตัวอย่าง

ตัวแปร	ขั้นที่ 1		ขั้นที่ 2	
	b	β	*b*	β
ค่าคงที่	8.361		14.896	
อายุ	-0.035	-.09	-0.030*	-.07
เพศชาย	0.500	.05	0.499	.05
ภาคเหนือ			-10.891**	-.88
ภาคกลาง			-9.898**	-.80
ภาคอีสาน			-6.139**	-.49
R^2	.01		.64**	
R^2 ที่เพิ่มขึ้น			.63**	

หมายเหตุ *p < .05, **p < .01

คาบต่อไป

- ทบทวนการวิเคราะห์ถดถอย
  - การวิเคราะห์ปฏิสัมพันธ์ (Interaction) หรือตัวแปรกำกับ (Moderation)