R PROGRAMMING Part 9







ผู้ช่วยศาสตราจารย์ ดร. อัชฌาณัท รัตนเลิศนุสรณ์

สาขาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี

สารบัญ

Contents

- การวิเคราะห์การถดถอยอย่างง่าย
- การวิเคราะห์สหสัมพันธ์อย่างง่าย

การวิเคราะห์การถดถอยอย่างง่าย

การวิเคราะห์การถดถอยอย่างง่ายเป็นการศึกษาถึงความสัมพันธ์ของตัว แปร 2 ตัว คือตัวแปรอิสระ (X) และตัวแปรตาม(Y) ในรูปแบบสมการ เชิงเส้นตรง อาทิ ต้องการหาสมการการถดถอยของตัวแปรดังต่อไปนี้

- ค่าโฆษณา(X) กับ ยอดขายสินค้า(Y)
- รายได้(X) กับ รายจ่าย(Y)
- ต้นทุน (X) กับราคาขาย(Y)

จุดประสงค์ของการวิเคราะห์การถดถอยอย่างง่าย

- 1. เพื่อหาตัวแบบหรือสมการแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ
- (X) กับตัวแปรตาม(Y) ในรูปแบบ

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
 , i = 1,2,3, ..., n

โดยที่ Y_i คือ ค่าสังเกตที่ i ของตัวแปรตาม

 X_i คือ ค่าสังเกตที่ i ของตัวแปรอิสระ

 \mathcal{E}_{i} คือ ค่าสังเกตที่ i ของความคลาดเคลื่อน

 $oldsymbol{eta_0}$ คือ ค่าคงที่ หรือ intercept จุดตัดแกน Y ของสมการการถดถอย

 $oldsymbol{eta_1}$ คือ ค่าสัมประสิทธิ์การถดถอยที่มีผลกับตัวแปรอิสระ(X) ของสมการการถดถอย

ซึ่งเป็นค่าที่แสดงถึงอัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย

2. เพื่อพยากรณ์หรือทำนายตัวแปรตาม เมื่อทราบค่าตัวแปรอิสระ

ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย

- 1. $arepsilon_i$ เป็นความคลาดเคลื่อนที่มีค่าเฉลี่ยเท่ากับศูนย์ หรือ $E(arepsilon_i)=0$
- 2. $arepsilon_i$ มีค่าความแปรปรวนเท่ากันทุกค่าของ i และมีค่าเท่ากับค่าแปรปรวน ของ Y_i นั่นคือ $V(arepsilon_i) = V(Y_i) = \sigma^2$
- 3. $arepsilon_i$ และ $arepsilon_j$ เป็นอิสระกัน นั่นคือ $COV\left(arepsilon_i,arepsilon_j
 ight)=0;\;i
 eq j$
- 4. $\varepsilon_i \sim normal(0, \sigma^2)$

ความหมายของค่าสัมประสิทธิ์ของการถดถอย

กรณีที่ 1 ถ้าค่า β₁> 0 แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางเดียวกัน นั่นคือ ถ้าค่าของ X เพิ่ม จะ ทำให้ค่าของ Y เพิ่มขึ้นด้วย แต่ถ้าค่าของ X ลดลง จะทำให้ค่าของ Y ลดลงด้วย

กรณีที่ 2 ถ้าค่า β₁< 0 แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางตรงข้ามกัน นั่นคือ ถ้าค่าของ X เพิ่ม จะทำให้ค่าของ Y ลดลง แต่ถ้าค่าของ X ลดลง จะทำให้ค่าของ Y เพิ่มขึ้น

<u>กรณีที่ 3</u> ถ้าค่า β_1 = 0 แสดงว่า X และ Y ไม่มีความสัมพันธ์กันเลย

สมการการถดถอยโดยประมาณ

เนื่องจากเราเก็บข้อมูลมาหาตัวแบบหรือสมการการถดถ[ื]อย ดังนั้นสมการที่ได้จึง เป็นสมการการถดถอยโดยประมาณซึ่งมีรูปแบบดังนี้

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
 , $i = 1, 2, 3, ..., n$

โดยที่ \widehat{Y}_i คือ ค่าประมาณที่ i ของตัวแปรตาม

 X_i คือ ค่าสังเกตที่ i ของตัวแปรอิสระ

 $\hat{oldsymbol{eta}}_{0}$ คือ ค่าประมาณของค่าคงที่ หรือ intercept จุดตัดแกน Y ของสมการการถดถอย

 $\hat{oldsymbol{eta}}_{1}$ คือ ค่าประมาณของสัมประสิทธิ์การถดถอยที่มีผลกับตัวแปรอิสระ(X) ของสมการการ ถดถอย

วิธีประมาณค่าสัมประสิทธิ์การถดถอย

วิธีประมาณค่าสัมประสิทธิ์ของการถดถอยที่นิยมใช้กัน คือ วิธีกำลัง สองน้อยที่สุด(Least square method) ทำให้ได้ค่าประมาณดังนี้

$$\hat{\beta}_{1} = \frac{\sum_{i=1}^{n} x_{i} y_{i} - \frac{\sum_{i=1}^{n} x_{i} \times \sum_{i=1}^{n} y_{i}}{n}}{\sum_{i=1}^{n} x_{i}^{2} - \frac{\left(\sum_{i=1}^{n} x_{i}\right)^{2}}{n}}$$

$$\hat{\beta}_{0} = \bar{Y} - \hat{\beta}_{1} \bar{X}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

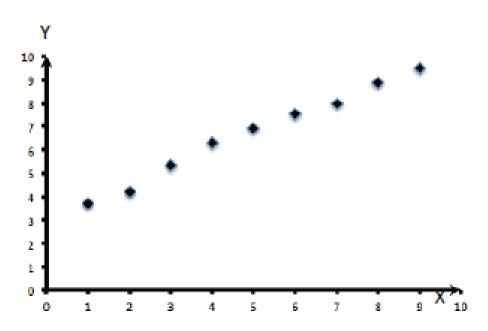
ตัวอย่างที่ 9.1 ในการศึกษาความสัมพันธ์ระหว่างค่าโฆษณากับยอดขายของบริษัทแห่งหนึ่ง เก็บข้อมูลย้อนหลัง 9 เดือน ได้ข้อมูลดังนี้

เดือนที่	1	2	3	4	5	6	7	8	9
ค่าโฆษณา(X) (หน่วย:แสนบาท)	1	2	3	4	5	6	7	8	9
ยอดขาย(Y) (หน่วย:ล้านบาท)	3.7	4.2	5.3	6.3	6.9	7.5	8.0	8.9	9.5

- 1) จงเขียนแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างค่าโฆษณากับยอดขาย
- 2) จงหาสมการการถดถอยระหว่างค่าโฆษณากับยอดขายของบริษัทนี้
- 3) จงพยากรณ์ยอดขายเดือนหน้า ถ้ากำหนดค่าโฆษณาในเดือนหน้าเป็น 550,000 บาท

วิธีทำ

ให้ Y = ยอดขาย และ X = ค่าโฆษณา
 นำข้อมูลที่มีอยู่มาเขียนกราฟ (แผนภาพการกระจาย) ได้ดังนี้



2) หาความสัมพันธ์ระหว่างค่าโฆษณากับยอดขายของบริษัทนี้

ยอดขาย(Y) (หน่วย:ล้านบาท)	ค่าโฆษณา(X) (หน่วย:แสนบาท)	x_i^2	$x_i y_i$	y_i^2		
3.7	1	1	3.7	13.69		
4.2	2	4	8.4	17.64		
5.3	3	9	15.9	28.09		
6.3	4	16	25.2	39.69		
6.9	5	25	34.5	47.61		
7.5	6	36	45	56.25		
8	7	49	56	64		
8.9	8	64	71.2	79.21		
9.5	9	81	85.5	90.25		
$\sum y = 60.3$	$\sum x = 45$	$\sum x^2 = 285$	$\sum xy = 345.4$	$\sum y_i^2 = 436.43$		

$$\hat{\beta}_{1} = \frac{\sum_{i=1}^{n} x_{i} y_{i} - \frac{\sum_{i=1}^{n} x_{i} \times \sum_{i=1}^{n} y_{i}}{n}}{\sum_{i=1}^{n} x_{i}^{2} - \frac{\left(\sum_{i=1}^{n} x_{i}\right)^{2}}{n}}$$

$$\hat{\beta}_1 = \frac{345.4 - \frac{45 \times 60.3}{9}}{285 - \frac{(45)^2}{9}} = \frac{43.9}{60} = 0.73$$

$$\overline{X} = \frac{45}{9}, \overline{Y} = \frac{60.3}{9}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_0 = \frac{60.3}{9} - 0.73 \frac{45}{9} = 3.05$$

ดังนั้น สมการการถดถอย คือ

$$\hat{Y}_i = 3.05 + 0.73X_i$$

ในที่นี้ $\hat{m{\beta}}_1$ = 0.73 หมายถึง ถ้าค่าโฆษณา (X) เพิ่มขึ้น 100,000 บาท (1 หน่วยของ X) ยอดขาย (Y) จะเพิ่มขึ้น 730,000 บาท (0.73 หน่วยของ Y)

3) จากสมการการถดถอย คือ

$$\hat{Y}_i = 3.05 + 0.73X_i$$

แทนค่า $X_i = 5.5$ ในสมการข้างต้นจะได้

$$\hat{Y}_i = 3.05 + 0.73x(5.5) = 7.065$$

นั่นคือ ถ้าให้ค่าโฆษณาในเดือนหน้าเป็น 550,000 บาท ยอดขายโดยประมาณในเดือนหน้าจะเป็น 7,065,000 บาท

สัมประสิทธิ์สหสัมพันธ์อย่างง่ายเป็นค่าสถิติที่ใช้วัดความสัมพันธ์ ระหว่างตัวแปร Y กับตัวแปร X ว่ามีความสัมพันธ์กันในเชิงเส้นตรง มากหรือน้อย โดยจะใช้สัญลักษณ์ ρ แทน ค่าสัมประสิทธิ์สหสัมพันธ์ เนื่องจากค่าสัมประสิทธิ์สหสัมพันธ์จะคำนวณจากข้อมูลตัวอย่างทำให้ ได้ค่าประมาณของสัมประสิทธิ์สหสัมพันธ์ ซึ่งแทนด้วย r ซึ่งมีค่าอยู่ ระหว่าง -1 ถึง 1 และไม่มีหน่วย

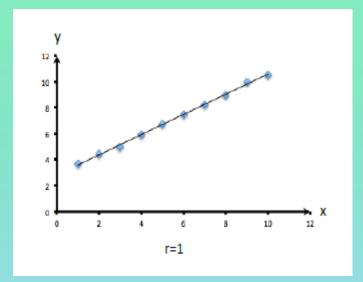
ความหมายของค่า $oldsymbol{r}$ สามารถแปลผลได้ดังนี้

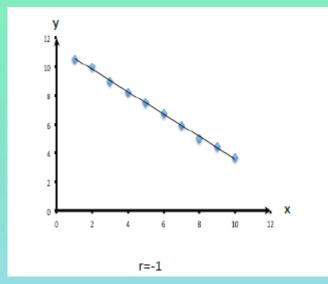
กรณีที่ 1 ถ้า r=0 แสดงว่า X และ Y ไม่มีความสัมพันธ์กัน

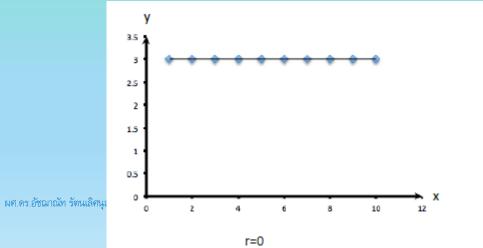
กรณีที่ 2 ถ้า $oldsymbol{\mathcal{T}}$ เข้าใกล้ 0 แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

กรณีที่ 3 ถ้า **?** เข้าใกล้ -1 แสดงว่า X และ Y มีความสัมพันธ์กันมากและมีความสัมพันธ์ใน ทิศทาง ตรงกันข้าม คือถ้า X เพิ่ม Y จะลด แต่ถ้า X ลด Y จะเพิ่ม

กรณีที่ 4 ถ้า *T* เข้าใกล้ 1 แสดงว่า X และ Y มีความสัมพันธ์กันมากและมีความสัมพันธ์ใน ทิศทาง เดียวกัน คือ ถ้า X เพิ่ม Y จะเพิ่มด้วยแต่ถ้า X ลด Y จะลดลงด้วย







ค่าสัมประสิทธิ์สหสัมพันธ์ $m{r}$ สามารถคำนวณได้จากสูตรดังนี้

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

ตัวอย่างการหาค่าสัมประสิทธิ์สหสัมพันธ์

ตัวอย่างที่ 9.2 จากการศึกษาความสัมพันธ์ระหว่างค่าโฆษณากับยอดขายของบริษัทแห่งหนึ่ง เก็บข้อมูลย้อนหลัง 9 เดือน ได้ข้อมูลดังนี้

เดือนที่	1	2	3	4	5	6	7	8	9
ค่าโฆษณา(X) (หน่วย:แสนบาท	1	2	3	4	5	6	7	8	9
ยอดขาย(Y) (หน่วย:ล้านบาท	3.7	4.2	5.3	6.3	6.9	7.5	8.0	8.9	9.5

จงหาสัมประสิทธิ์สหสัมพันธ์ของค่าโฆษณากับยอดขาย พร้อมทั้งอธิบายความหมาย

ตัวอย่างการหาค่าสัมประสิทธิ์สหสัมพันธ์

วิธีทำ จากตัวอย่างที่ 9.1 ได้ $S_{XX}=60,\ S_{XY}=43.9$

พิจาชณาหา
$$S_{YY} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 436.43 - \frac{(60.3)^2}{9} = 32.42$$

แทนค่าลงในสมการ (9.6) จะได้

$$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{43.9}{\sqrt{60(32.42)}} = 0.9954$$

r=0.9954 หมายความว่า ค่าโฆษณากับยอดขายมีความสัมพันธ์กันมากและเป็นไปใน

ทิศทาง เดียวกัน

สัมประสิทธิ์การตัดสินใจ

(Coefficient of Determination : R^2 หรือ r^2)

สัมประสิทธิ์การตัดสินใจ เป็นค่าที่ใช้บอกว่าตัวแปร X สามารถอธิบายการ เปลี่ยนแปลงของตัวแปร Y ได้มากน้อยเท่าไหร่

โดยใช้สัญลักษณ์ R^2 แทน ค่าพารามิเตอร์ของสัมประสิทธิ์การตัดสินใจ และ \boldsymbol{r}^2 แทน ค่าประมาณสัมประสิทธิ์การตัดสินใจจากตัวอย่าง

โดยที่ $0 \leq r^2 \leq 1$

ถ้ามีค่ามาก(เข้าใกล้ 1)แสดงว่าตัวแปร X สามารถอธิบายการเปลี่ยนแปลงของค่า Y ได้มาก

ถ้ามีค่ามาก(เข้าใกล้ 0)แสดงว่าตัวแปร X สามารถอธิบายการเปลี่ยนแปลงของค่า Y ได้น้อยมาก

การคำนวณค่าสัมประสิทธิ์การตัดสินใจ

สัมประสิทธิ์การตัดสินใจ สามารถคำนวณได้จากสูตรดังนี้

$$r^2 = b \frac{ss_{xy}}{ss_{yy}} = \frac{\hat{\beta}_1 s_{xy}}{s_{yy}}$$

โดยที่
$$S_{xx} = \sum (x_i - \overline{x})^2$$
 = $\sum x_i^2 - \frac{(\sum x_i)^2}{n}$ $S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$ = $\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$ $S_{YY} = \sum (y_i - \overline{y})^2$ = $\sum y_i^2 - \frac{(\sum y_i)^2}{n}$ $\overline{x} = \frac{\sum x_i}{n}$, $\overline{y} = \frac{\sum y_i}{n}$

เมื่อ S_{XX} คือ ความแปรปรวนของ X ทั้งหมด S_{YY} คือ ความแปรปรวนของ Y ทั้งหมด S_{XY} คือ ความแปรปรวนของ Y ที่เกิดจาก X

ตัวอย่างการหาค่าสัมประสิทธิ์การตัดสินใจ

ตัวอย่างที่ 9.3 จากตัวอย่างที่ 9.1 จงหาสัมประสิทธิ์การตัดสินใจ พร้อมทั้งอธิบายความหมาย

วิธีทำ จากตัวอย่างที่ 9.1 และ 9.2 ได้ $S_{XY}=43.9,\ S_{YY}=32.42$, b=0.73 แทนค่าลงในสมการ (9.7) จะได้

$$\therefore r^2 = b \frac{SS_{xy}}{SS_{yy}} = 0.73 \frac{43.9}{32.42} = 0.9885$$
 หรือ 98.85%

 $r^2=0.9885$ หมายความว่า ค่าโฆษณาสามารถอธิบายการเปลี่ยนแปลงของยอดขายได้ 98.85% ส่วนที่เหลือเกิดจากปัจจัยอื่น

สรุปขั้นตอนการวิเคราะห์การถดถอยและสหสัมพันธ์อย่างง่าย

- พิจารณาความสัมพันธ์ของตัวแปรอิสระ X กับตัวแปรตาม Y จากการเขียนแผนภาพการ กระจาย
 - 2. ถ้า X และ Y มีความสัมพันธ์กันในรูปเส้นตรง จะได้สมการที่แสดงความสัมพันธ์คือ

$$Y = \beta_0 + \beta_1 X + e$$

- 3. ใช้วิธีกำลังสองน้อยที่สุดในการประมาณค่า eta_0 และ eta_1 ด้วย a และ b ตามลำดับ ซึ่งจะได้ ค่าประมาณจากสมการ \widehat{Y} = a + bx
 - 4. ใช้สมการที่ได้ในขั้นตอนที่ 3 ประมาณค่า Y เมื่อกำหนดค่า X
 - 5. คำนวณหาสัมประสิทธิ์สหสัมพันธ์ตัวอย่าง (r) หรือหาสัมประสิทธิ์การตัดสินใจ (r^2)

lm {stats}

Fitting Linear Models

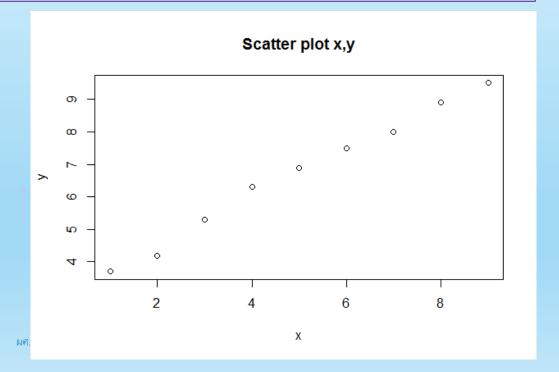
Description

lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
    method = "qr", model = TRUE, x = FALSE, y = FALSE,
    qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset,
...)
```

```
> #ex.9.1
> x=1:9
> y=c(3.7,4.2,5.3,6.3,6.9,7.5,8.0,8.9,9.5)
> #1. scatter plot(x,y)
> plot(x,y,type="p",main="Scatter plot x,y")
```



```
> #2. fit simple regression model
> lm(y\sim x)
call:
lm(formula = y \sim x)
Coefficients:
(Intercept)
                    0.7317
     3.0417
```

สมการการถดถอยโดยประมาณ คือ $\hat{Y}_i = 3.0417 + 0.7317X_i$

```
> #3. x=5.5 , yhat=?
> yhat=3.0417+0.7317*(5.5)
> print(yhat)
[1] 7.06605
```

แปลความหมายได้ว่าเมื่อกำหนดค่าโฆษณาเท่ากับ 5.5 (หน่วย: แสนบาท) จะพยากรณ์ยอดขายได้เท่ากับ 7.06605 (หน่วย: ล้านบาท)

```
> #4. R^2
> x=1:9
> y=c(3.7,4.2,5.3,6.3,6.9,7.5,8.0,8.9,9.5)
> ols.model=lm(y~x)
> summary(ols.model)
call:
lm(formula = y \sim x)
Residuals:
    Min
         1Q Median
                              3Q
                                      Max
-0.30500 -0.12667  0.00500  0.06833  0.33167
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept 0.15035 20.23 1.81e-07 ***
         R 0.73167 0.02672 27.38 2.22e-08 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.207 on 7 degrees of freedom
Multiple R-squared: 0.9908, Adjusted R-squared: 0.9894
F-statistic: 749.9 on 1 and 🗙 DF, p-value: 2.222e-08
```

Multiple R-squared: 0.9908 แปลความหมายได้ว่าค่าโฆษณา(ตัวแปร อิสระ)สามารถอธิบายการเปลี่ยนแปลงของค่ายอดขาย(ตัวแปรตาม) ได้สูง ถึง 99.08 เปอร์เซ็นต์

```
> #5. r=correlation of x,y
> x=1:9
> y=c(3.7,4.2,5.3,6.3,6.9,7.5,8.0,8.9,9.5)
> cor(x,y)
[1] 0.9953651
```

r = 0.9953651 แปลความหมายได้ว่าค่าโฆษณา(ตัวแปรอิสระ)กับ ยอดขาย(ตัวแปรตาม) มีความสัมพันธ์กันสูงและมีทิศทางเดียวกัน

ฝึกปฏิบัติการตามตัวอย่างที่ 9.4 ด้วยโปรแกรมอาร์

ตัวอย่างที่ 9.4 จากการศึกษารายได้กับรายจ่ายของประชาชนในเขตหนึ่ง มีข้อมูลดังนี้

คนที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
รายได้ (X) (หน่วย:หมื่นบาท)	0.8	0.9	1	1.2	1.3	1.5	1.9	2	2.2	2.3	2.4	2.5	3	3.5	4
รายจ่าย (Y) (หน่วย:หมื่นบาท)	0.5	0.6	0.9	1.4	1.6	1.4	1.8	2	2.1	2.2	2.3	2.5	2.7	3.4	3.7

- 1) จงเขียนแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างรายได้กับรายจ่าย
- 2) จงหาสมการถดถอยของความสัมพันธ์ระหว่างรายได้กับรายจ่าย
- 3) จงพยากรณ์รายจ่าย ถ้ากำหนดค่ารายได้ของประชนเป็น 14,000 บาท
- 4) จงหาสัมประสิทธิ์สหสัมพันธ์ พร้อมทั้งอธิบายความหมาย
- 5) จงหาสัมประสิทธิ์การตัดสินใจ พร้อมทั้งอธิบายความหมาย

