# บทที่ ๓

## โมเดลที่ใช้ในการพยากรณ์

ในที่นี้จะทำการศึกษาเปรียบเทียบจากผลของการวิเคราะห์อนุกรมเวลา(Time Series Analysis) กับการวิเคราะห์ความถดถอย (Regression Analysis)ในการพยากรณ์ปริมาณการบริโภคและปริมาณผลผลิตน้ำตาลทรายภายในประเทศ เพื่อที่จะได้เลือกวิธีใดที่ให้ผลในการพยากรณ์ได้ดีกว่ากัน

๓.๑ การวิเคราะห์อนุกรมเวลา (Time Series Analysis)

การวิเคราะห์อนุกรมเวลา คือการศึกษาถึงความเคลื่อนไหวของข้อมูล ชุดหนึ่ง ๆ ตามงวดระยะเวลา ข้อมูลต่าง ๆ อาทิ เช่น ข้อมูลของปริมาณการบริโภคน้ำตาลทรายภายในประเทศ และข้อมูลของปริมาณผลผลิตน้ำตาลทราย ภายในประเทศที่เก็บรวบรวมมานั้น เป็นข้อมูลที่ได้จัดให้เรียงลำดับตามเวลา (ตามปี พ.ศ.) ที่เกิดขึ้นมาแล้ว

ส่วนประกอบของอนุกรมเวลา มีดังนี้ :-

๓.๑.๑ แนวโน้มตามลำดับเวลา (Secular Trend หรือ Trend หรือ Secular Movements) ซึ่งจะใช้อักษรย่อว่า "T" เป็นการเคลื่อนไหวของข้อมูลในระยะเวลาที่ค่อนข้างจะยาว ตามปกติจะแสดงถึงทิศทางของอนุกรมเวลาชุดนั้น ๆ พุ่งไปสู่ ซึ่งอาจจะมีลักษณะ เป็นเส้นตรง (Linear Trend) หรือ เส้นโค้ง (Non Linear Trend ) ก็ได้ หน่วยของระยะเวลาตามปกติเป็นปี

๓.๑.๒ การเปลี่ยนแปลงตามฤดูกาล (Seasonal Variations หรือ Seasonal movements) จะใช้อักษรย่อว่า "S" เป็นการเปลี่ยนแปลงของข้อมูลที่มักจะมีลักษณะเปลี่ยนแปลงขึ้น ๆ ลง ๆ อย่างเดียวกันหรือ คล้าย ๆ กัน ในช่วงระยะเวลาที่กำหนด หน่วยของระยะเวลาอาจเป็นงวด ๓ เดือน รายเดือน รายสัปดาห์ หรือรายวันก็ได้

๓.๑.๓ การเปลี่ยนแปลงตามวัฏจักร (Cyclical Variations หรือ

Cyclical Movements ) จะใช้อักษรย่อ "C" เป็นการเปลี่ยนแปลงของข้อมูล ที่มีลักษณะคล้ายลูกคลื่น กล่าวคือจะมีลักษณะของข้อมูลพุ่งสูงขึ้น ชั่วระยะเวลาหนึ่ง และติดตามด้วยการลดต่ำลงในอีกช่วงระยะหนึ่งสลับกันไป

๓.๑.๔ การเปลี่ยนแปลงเนื่องจากเหตุการณ์ผิดปกติหรือเป็นเชิงสุ่ม (Irregulation Variations หรือ . Irregulation Movements หรือ Random Movements ) จะใช้อักษรย่อว่า "I" เป็นการเปลี่ยนแปลงที่ไม่อาจคาดการณ์ได้ ล่วงหน้า เหตุการณ์ผิดปกตินี้อาจเป็นเรื่องเล็ก ๆ น้อย ๆ เช่นการนัดหยุดงานของคนงานไปจนถึงเรื่องใหญ่ ๆ เช่น การเกิดแผ่นดินไหว หรือการเกิดสงคราม ซึ่งเป็นปัจจัยที่มิได้อยู่ภายใต้กฎเกณฑ์ที่แน่นอนแต่อย่างใด

ในอนุกรมเวลาชุดหนึ่ง ๆ อาจมีส่วนประกอบเฉพาะบางส่วน หรือมีรวมกันทั้งหมด ( T, S, C, I ) ก็เป็นได้ และโมเดล ( Model ) ขั้นมูลฐานที่ใช้ในการแสดงส่วนประกอบของอนุกรมเวลามี ๒ โมเดล

ก. โมเดลเชิงบวก ( Additive Model )

ได้แก่การนำเอาองค์ประกอบทั้ง ๔ มาบวกกันจะได้ Y = T+ S+ C+ I

ข. โมเดลเชิงคูณ (Multiplicative Model )

ได้แก่การนำเอาองค์ประกอบทั้ง ๔ มาคูณกันจะได้ Y = T.S.C.I

โมเดลที่จะใช้ในที่นี้เป็นโมเดลเชิงคูณ

๓.๒ วิธีการกะประมาณค่าแนวโน้มตามลำดับเวลา (หรือวิธีคำนวณหาเส้น แนวโน้ม)

ปกติในการวิเคราะห์ แนวโน้มตามลำดับเวลา มักจะใช้ข้อมูลรายปีแทนที่จะเป็นงวด ๓ เดือน รายเดือน หรือสัปดาห์ ซึ่งจะทำให้เกิดความยุ่งยากในการคำนวณ โดยไม่จำเป็น โดยนำข้อมูลที่เตรียมไว้แล้วมาเขียนร่างในกระดาษกราฟก่อน เพื่อจะได้พิจารณาลักษณะของ แนวโน้มที่ควรจะเป็น ว่าจะมีลักษณะเป็นเส้นตรงหรือ เส้นโค้ง แล้วจะได้นำเอาเส้นแนวโน้มนั้น ๆ เพื่อใช้ในการพยากรณ์ต่อไป วิธีการกะประมาณค่าแนวโน้มตามลำดับเวลา มีอยู่หลายวิธีด้วยกัน แต่วิธีที่เห็นว่าเหมาะสม นิยมใช้มากที่สุด และ

ง่ายในการคำนวณ โดยใช้หลักที่ว่า ผลรวมของผลต่างระหว่างค่าของข้อมูลที่ได้ จากเส้นแนวโน้มกับค่าจริงของข้อมูล ยกกำลังสองแล้วได้ค่าน้อยที่สุด ซึ่งเรียกว่าวิธีกำลังสองน้อยที่สุด (Least Squares Method)

๓.๖.๑ กรณีที่ได้แนวโน้มมีลักษณะเป็นเส้นตรง (Linear Trend)

โมเดลที่ใช้แทนเส้นตรงคือ

$$Y = A + BX + E \quad \text{.....} \quad (1)$$

A ค่าคงที่หรือเป็นค่า Y เมื่อ X = 0 (ระยะตัดแกน Y )

B ค่าความชัน (Slope ) ของเส้นตรง

E ค่าผิดพลาด (Error หรือ Disturbance Term )

X เป็นตัวแปรอิสระ (ในที่นี้คือระยะเวลา มีหน่วยเป็น ปี)

Y เป็นค่าจริงของข้อมูล (เป็นตัวแปรตาม)

ค่าประมาณของ (1) คือ

$$\hat{Y} = a + bX \quad \text{.....} \quad (2)$$

$\hat{Y}$ คือค่าประมาณของ Y

a คือค่าประมาณของ A

b คือค่าประมาณของ B

$e = Y - \hat{Y}$ = ความแตกต่างระหว่างค่าจริงกับค่าประมาณหรือความคลาดเคลื่อนหรือ Residual

$$\sum e^2 = \sum (Y - \hat{Y})^2 \quad \text{.....} \quad (3)$$

ในการคำนวณหาค่า a, b ซึ่งเป็นค่าคงที่ เพื่อให้ได้ความคลาดเคลื่อนน้อยที่สุดจะใช้วิธีกำลังสองน้อยที่สุด ก็จะได้ a, b เป็นค่า Unbiased Estimate และมี Standard Error น้อยที่สุดด้วยเมื่อเปรียบเทียบกับ Unbiased Estimate ตัวอื่น ๆ ที่คำนวณได้จากโมเดลเดียวกัน

การหาค่า a,b ทำได้โดยใช้หาเชียล เดริเวทีฟ (Partial derivative) ของ (3) เทียบกับ a , b แล้วให้มีค่าเป็น 0

$$\frac{\partial \Sigma e^2}{\partial a} = 0 \quad , \quad \frac{\partial \Sigma e^2}{\partial b} = 0$$

จะได้ Normal Equation ดังนี้

$$\left.\begin{aligned} \Sigma Y &= Na + b \Sigma X \\ \Sigma XY &= a\Sigma X + b \Sigma X^2 \end{aligned}\right\} \quad \cdots\cdots\cdots\cdots\cdots (1)'$$

เนื่องจากข้อมูลมี ๑๔ ปีซึ่งเป็นเลขคู่ เพื่อง่ายแก่การคำนวณจึงเลือกใช้ปีดังนี้

ให้ปี ๒๕๐๕ มีค่า X = - 13

๒๕๐๖ มีค่า X = - 11

๒๕๐๗ มีค่า X = - 9

. .

. .

. .

. .

. .

๒๕๑๘ มีค่า X = 13

ดังนั้น $\Sigma X = 0$ , $\Sigma X^3 = 0$

และ $\Sigma X^2 = 910$ , $\Sigma X^4 = 105,742$

เพื่อความสดวกในการคำนวณ จาก (1)' เขียนได้ใหม่เป็น

$\Sigma Y = 14a$ , $N = 14$

$\Sigma XY = 910b$

หรือ $a = \Sigma Y / 14$

$b = \Sigma XY / 910$ ..... (2)'

๓.๒.๒ กรณีที่ได้แนวโน้มมีลักษณะเป็นเส้นโค้ง ( Non Linear Trend ) กรณีที่กาดว่าลักษณะข้อมูลที่ได้มีลักษณะเป็นเส้นโค้ง เมื่อทดลองนำไปเขียนโดยพิจารณา จาก Scatter Diagram การเลือกลักษณะโมเดลโดยทั่ว ๆ ไป จะอยู่ใน ๒ ลักษณะดังนี้คือ

ก. โมเดลที่อยู่ในรูปโพลี่โนเมี่ยลกำลังสอง (Quadratic Polynomial )

$$\hat{Y} = a + bX + cX^2 \quad \text{.....} \quad (4)$$

โดยที่ a เป็นค่าของ $\hat{Y}$ เมื่อ X มีค่า = 0

b เป็นค่าของความชัน (Slope ) ของแนวโน้ม

c เป็นอัตราการเปลี่ยนของความชัน

การหาค่า a, b , c ทำโดยวิธีใช้กำลังสองน้อยที่สุดเช่นกัน จะได้ Normal Equation ดังนี้

$$\left.\begin{aligned} \Sigma Y &= Na + b\Sigma X + c\Sigma X^2 \\ \Sigma XY &= a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \\ \Sigma X^2Y &= a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 \end{aligned}\right\} \quad \text{.....} \quad (4)'$$

แทนค่า $N, \Sigma X, \Sigma X^2, \Sigma X^3, \Sigma X^4$ จะได้

$$\left.\begin{aligned} \Sigma Y &= 14a + 910c \\ \Sigma XY &= 910b \\ \Sigma X^2Y &= 910a + 105,742c \end{aligned}\right\} \text{-----} (5)'$$

ดังนั้น

$$\left.\begin{aligned} a &= \frac{105,742(\Sigma Y) - 910(\Sigma X^2Y)}{14(105,742) - (910)^2} \\ b &= \Sigma XY / 910 \\ c &= \frac{14(\Sigma X^2Y) - 910(\Sigma XY)}{14(105,742) - (910)^2} \end{aligned}\right\} \text{--} (6)'$$

พิจารณา $\dfrac{14(105,742) - (910)^2}{14} = 105,742 - 59,150$

$= 46,592$

$\dfrac{105,742}{14} = 7,553$

$910/14 = 65$

$7,553 / 46,592 = 0.162109$

$65 / 46,592 = 0.001395$

$1 / 46,592 = 0.0000214$

ดังนั้นจาก (6)' จะได้

$$\left.\begin{aligned} a &= 0.162109\Sigma Y - 0.001395\ \Sigma X^2Y \\ b &= 0.001099\ \Sigma XY \\ c &= 0.0000214\ \Sigma X^2Y - 0.001395\ \Sigma Y \end{aligned}\right\} \text{--} (7)'$$

ข. โมเดลที่อยู่ในรูปเอกโพเนนเชียล (Exponential Trend)

มีรูป $Y = ab^X$ .....(5)

a คือค่า $\hat{Y}$ เมื่อ $X = 0$

b อัตราการเพิ่ม (rate of increase)

b ถ้ามีค่าอยู่ระหว่าง 0 กับ 1 แสดงว่า Y จะมีค่าลดลงเมื่อ X มีค่าเพิ่มขึ้น

b ถ้ามากกว่า 1 แสดงว่า Y จะมีค่าเพิ่ม เมื่อ X มีค่าเพิ่มขึ้น

ในการหาค่า a,b จะยุ่งยาก วิธีที่สดวกทำโดยการ Take ใส่ log

สมการ (5) จะได้

$$\log \hat{Y} = \log a + X \log b \quad \text{--------(6)}$$

การหาค่า log a, log b จะเหมือนกรณีเส้นตรง โดยใช้วิธีกำลังสองน้อยที่สุดเช่นกันจะได้ Normal Equation ดังนี้

$$\left.\begin{aligned} \Sigma \log y &= N \log a + (\log b)(\Sigma X) \\ \Sigma X \log Y &= (\log a)(\Sigma X) + (\log b)(\Sigma X^2) \end{aligned}\right\} \text{--(7)}$$

แทนค่า $N, \Sigma X, \Sigma X^2$ จะได้

$$\Sigma \log Y = 14 \log a$$

$$\Sigma \log Y = 910 \log b$$

ดังนั้น

$$\left.\begin{aligned} \log a &= (\Sigma \log Y) / 14 \\ \log b &= (\Sigma X \log Y) / 910 \end{aligned}\right\} \text{--------(8)}$$

เมื่อเปิดค่า Anti log จะได้ค่า a,b ตามต้องการ

การพิจารณาดูจาก Scatter Diagram จะไม่สามารถบอกรายละเอียด ได้มากนัก จึงเห็นสมควรทำการทดสอบค่าลักษณะของเส้นแนวโน้มที่ควรจะเป็นใน ๓ โมเดล คือ :-

๑. โมเดลของเส้นตรง

๒. โมเดลของโพลี่โนเมี่ยลกำลังสอง

๓. โมเดลของเอ็กซโพเนนเชี่ยล

จากนั้นจึงพิจารณาถึง ผลบวกกำลังสองของผลต่างระหว่างข้อมูลเดิมกับที่คำนวณได้จากเส้นแนวโน้ม (Sum of Square)ในแต่ละโมเดล เพื่อดูว่าค่าที่คำนวณได้จากโมเดลใด มีค่าน้อยที่สุดก็เลือกเอาโมเดลนั้นเป็นตัวแทนของ Time Series Analysis หรือการเลือกโมเดล โดยการคำนวณหาค่า Standard Error of Estimate of Y on X or Standard Diviation of Regression

$$S_{Y.X} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{(n - k)}} \qquad \text{.....} (9)$$

n = จำนวน ข้อมูล = 14

k = จำนวนพารามีเตอร์ = 2 (สำหรับเส้นตรง)

และคำนวณหาค่าเปอร์เซนต์ของการกระจายของ Y (Dependent Variable) ซึ่งคำนวณได้จาก

$$\frac{\Sigma (\hat{Y} - \bar{Y})^2}{\Sigma (Y - \bar{Y})^2} \times 100 \qquad \text{.....} (10)$$

$\bar{Y}$ คือค่ามัชฌิมเลขคณิตของตัวแปร ไม่อิสระ

โมเดลใดที่ให้ค่า S ต่ำสุดและเปอร์เซนต์ของการกระจายมากที่สุดก็เลือกใช้โมเดลนั้นเป็นตัวแทน แต่บางครั้งก็อาจมีปัญหาเกิดขึ้น กล่าวคือเมื่อได้ทำการเลือกตัดสิน ตามกฎเกณฑ์ดังกล่าวแล้ว เมื่อนำไปใช้พยากรณ์ในอนาคตจะมีแนวโน้มที่ Diverge ค่อนข้างเร็ว ทำให้ค่าในอนาคตเปลี่ยนแปลงเร็วเกินความเป็นจริงบางครั้งจำเป็นต้องเลือกโมเดลอื่นที่เหมาะและสามารถให้เหตุผลพอประมาณในการพยากรณ์ได้

๓.๒.๓ วิธีการทดสอบโมเดล

การทดสอบความมีนัยสำคัญของโมเดล โดยใช้ **F - test** เมื่อเลือกโมเดลได้ตามลักษณะของการพิจารณาดังกล่าว จำเป็นที่จะต้องนำโมเดลที่ได้มาทดสอบหาความมีนัยสำคัญ ภายใต้สมมติฐาน

$$Ho : B = 0 \quad \text{หรือ} \quad B = C = 0$$

ค่าของ F พิจารณาได้จากตารางการวิเคราะห์ความแปรปรวน (Analysis of Variance ) หรือตาราง ANOVA

**ANOVA** [๑]

| Source of Variation | Degree of Freedom | Sum of Square | Mean Square | $F_c$ |
|---|---|---|---|---|
| Regression | k | $\sum (\hat{Y} - \bar{Y})^2 = SSR$ | $MSR = SSR/k$ | $\frac{MSR}{MSE}$ |
| Residual | n-k-1 | $\sum (Y - \hat{Y})^2 = SSE$ | $MSE = SSE/(n-k-1)$ | |
| Total | n -1 | $\sum (Y - \bar{Y})^2 = SST$ | | |

ต่อไปพิจารณาค่า $F_c$ ที่คำนวณได้จาก ANOVA เปรียบเทียบกับค่า $F_t$ จากตาราง F - distribution ที่ d.f = (k,n-k-1) ณ ระดับความเชื่อมั่นต่าง ๆ อาทิเช่น $\alpha = 0.01$ หรือ 0.05 ถ้าปรากฏว่า ค่า $F_c > F_t$ จะหมายความว่าเราจะปฏิเสธสมมติฐานที่ตั้งขึ้น นั่นคือเราจะยอมรับว่าโมเดลที่นำมาทำการทดสอบนั้นใช้ได้

---

๑ STEEL AND TORRIE , PRINCIPLES AND PROCEDURE OF STATISTICS (Mc. GRAW - HILL ) P 296

หมายเหตุ ค่าของ $\sum (Y - \bar{Y})^2$ เรียกว่า Total Variation

ค่าของ $\sum (Y - \hat{Y})^2$ เรียกว่า Unexplained Variation

ค่าของ $\sum (\hat{Y} - \bar{Y})^2$ เรียกว่า Explained Variation

$$\text{Coefficient of determination:}(R^2) = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

๓.๓ การวิเคราะห์ความถดถอย (Regression Analysis)

เนื่องจากในการวิเคราะห์ปริมาณการบริโภคน้ำตาลและปริมาณผลผลิตน้ำตาลทรายภายในประเทศมีลักษณะเกี่ยวข้องกับหลายตัวแปร จึงจำเป็นที่จะต้องใช้การวิเคราะห์ความถดถอยเชิงซ้อน (Multiple Linear Regression ) ซึ่งเป็นการวิเคราะห์ความสัมพันธ์ของข้อมูลตั้งแต่ ๒ ชุดขึ้นไป คือเมื่อกำหนดให้ข้อมูลชุดหนึ่งเป็นตัวแปรตาม และข้อมูลชุดอื่น ๆ เป็นตัวแปรอิสระ ในรูปสมการกำลังหนึ่ง

$$Y = B_o + B_1 X_1 + B_2 X_2 + B_3 X_3 + \cdots\cdots + B_k X_k + E$$

ค่าประมาณของสมการนี้คือ $\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k \cdots\cdots (1)$

$X_i$ เป็นตัวแปรอิสระที่ i โดยที่ i = 1,2,3 .....k

Y เป็นตัวแปรตาม

$B_i$ เป็นค่าสัมประสิทธิ์ของตัวแปรอิสระตัวที่ i, i = 1,2 3.....k

$\hat{Y}$ เป็นค่าประมาณของ Y

$b_i$ เป็นค่าประมาณของ $B_i$

a เป็นค่าประมาณของ $B_o$

๓.๓.๑ การคำนวณหาค่าสัมประสิทธิ์ตัวแปรอิสระในสมการความถดถอยเชิงซ้อน

จากสมการ (A) ให้ Y เป็นค่าจริงของข้อมูล

$\hat{Y}$ เป็นค่าประมาณของข้อมูลใน (A)

e = ความคลาดเคลื่อน $= Y - \hat{Y}$

= ความแตกต่างระหว่างค่าจริงกับค่าประมาณของข้อมูล

$e = Y - a - b_1X_1 - b_2X_2 - \cdots\cdots\cdots - b_kX_k$

ถ้ามี n ข้อมูลจะได้

$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (Y_i - a - b_1X_1 - \cdots\cdots\cdots - b_kX_k)^2 \cdots\cdots\cdots (B)$$

โดยวิธีกำลังสองน้อยที่สุดจะได้ค่า a, $b_1$, $b_2$, ... $b_k$ เป็นค่า Unbiased Estimate และมีค่า Standard Error น้อยที่สุดเมื่อเปรียบเทียบกับ Unbiased Estimate ตัวอื่น ๆ ที่คำนวณได้จากโมเดลเดียวกัน

<u>การคำนวณหาค่า a</u>

โดยการใช้พาเชี่ยลเดริเวทีฟ (Partial derivative) ของ (B) เทียบกับ a แล้วให้มีค่า = 0 นั่นคือ

$$\left( \frac{\partial \sum e^2}{\partial a} \right) = 0$$

$$-2 \sum (Y - a - b_1X_1 - b_2X_2 - \cdots\cdots\cdots - b_kX_k) = 0$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \cdots\cdots\cdots - b_k\bar{X}_k$$

แทนค่า a ใน (B) จะได้

$$\Sigma e^2 = \Sigma \left\{ Y - \bar{Y} - b_1 (X_1 - \bar{X}_1) \ldots\ldots b_k (X_k - \bar{X}_k) \right\}^2$$

ให้ $w_i = Y_i - \bar{Y}_i$

$x_i = X_i - \bar{X}_i$

$$\therefore \quad \Sigma e^2 = \Sigma (y - b_1 x_1 - b_2 x_2 \ldots\ldots\ldots\ldots b_k x_k)^2 \quad \text{---} \quad (C)$$

การคำนวณหาค่า $b_i$

โดยการใช้พาเ็ยลเดริเวทีฟ (Partial derivative ) ของ (B) เทียบกับ $b_i$ แล้วให้มีค่า =0 จะได้ k สมการซึ่งเป็น **Normal Equation** สำหรับ $X_i$ k ตัว

$$\left.\begin{aligned} \frac{\partial \Sigma e^2}{\partial b_1} = 0 \; ; \; b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + \ldots\ldots + b_k \Sigma x_1 x_k &= \Sigma x_1 y \\ b_1 \Sigma x_2 x_1 + b_2 \Sigma x_2^2 + \ldots\ldots + b_k \Sigma x_2 x_k &= \Sigma x_2 y \\ \vdots \\ \frac{\partial \Sigma e^2}{\partial b_k} = 0 \; ; \; b_1 \Sigma x_k x_1 + b_2 \Sigma x_k x_2 + \ldots\ldots + b_k \Sigma x_k^2 &= \Sigma x_k y \end{aligned}\right\} \quad (D)$$

จาก (D) แก้สมการหาค่า $b_1, b_2 \ldots b_k$ ได้จาก k สมการ และกรณีที่ข้อมูลมี $X_i$ มากกว่า $n$ ตัว ขึ้นไปควรจะหาค่า $b_i$ โดยวิธีเมทริกซ์ (Matrix) จะสดวกกว่าโดยวิธี Matrix

สมมติมี n ข้อมูลและมีตัวแปรอิสระ $n$ ตัวคือ k = 3 เขียนเรียงลำดับได้ในรูป Matrix ดังนี้ :-

| ลำดับ | Y | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| 1 | $Y_1$ | $X_{11}$ | $X_{21}$ | $X_{31}$ |
| 2 | $Y_2$ | $X_{12}$ | $X_{22}$ | $X_{32}$ |
| 3 | $Y_3$ | $X_{13}$ | $X_{23}$ | $X_{33}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| n | $Y_n$ | $X_{1n}$ | $X_{2n}$ | $X_{3n}$ |

หรือเขียนใหม่ได้เป็น

$$Y = X\beta + \epsilon$$

โดยให้ $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}$ , $X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ \cdot & & \\ \cdot & & \\ X_{n1} & X_{n2} & X_{n3} \end{bmatrix}$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \epsilon = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

คำนวณหาค่า

$$X'X = \begin{bmatrix} \Sigma X_{i1}^2 & \Sigma X_{i1}X_{i2} & \Sigma X_{i1}X_{i3} \\ \Sigma X_{i2}X_{i1} & \Sigma X_{i2}^2 & \Sigma X_{i2}Xi3 \\ \Sigma X_{i3}X_{i1} & \Sigma X_{i3}X_{i2} & \Sigma X_{i3}^2 \end{bmatrix}$$

และได้ $$X'Y = \begin{bmatrix} \Sigma X_{i1}Y_i \\ \Sigma X_{i2}Y_i \\ \Sigma X_{i3}Y_i \end{bmatrix}$$

ได้ Normal Equation

$$(X'X)\hat{\beta} = X'Y$$

$$(X'X)^{-1}\ (X'X)\hat{\beta} = (X'X)^{-1}\ (X'Y)$$

$$\hat{\beta} = (X'X)^{-1}\ (X'Y)$$

$$\hat{\beta} = C\ (X'Y)$$

ให้ $C = (X'X)^{-1} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$

จะได้

$$\beta_1 = C_{11} \sum X_{i1}Y_i + C_{12} \sum X_{i2} Y_i + C_{13} \sum X_{i3} Y_i$$
$$\beta_2 = C_{21} \sum X_{i1} Y_i + C_{22} \sum X_{i2} Y_i + C_{23} \sum X_{i3} Y_i$$
$$\beta_3 = C_{31} \sum X_{i1} Y_i + C_{32} \sum X_{i2} Y_i + C_{33} \sum X_{i3} Y_i$$

ถ้ากำหนดให้ $\hat{y} = \hat{Y} - \bar{Y}$

$$e = Y - \hat{Y}$$

$$y = Y - \bar{Y}$$
$$= (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$
$$= \hat{y} + e$$

ดังนั้น $\sum y^2 = \sum \hat{y}^2 + \sum e^2$

จากสมการข้างบนจะเห็นว่า ผลบวกกำลังสองของ deviation ของ Y จาก mean ประกอบด้วย ๒ ส่วนดังนี้

ก. $\sum \hat{y}^2$ เรียกว่า Sum of Squares เนื่องมาจากความถดถอยซึ่งเป็นผลบวกกำลังสองของ deviation $\hat{Y}$ จาก mean

ข. $\sum e^2$ เป็นผลบวกกำลังสองของ deviation ของค่าจริงจากค่าประมาณ

การคำนวณหาค่า F value เพื่อทดสอบโมเดลสมการถดถอยเชิงซ้อนได้จากตาราง ANOVA สำหรับ Multiple Linear Regression

ANOVA

| Source of Variation | d.f | Sum of Squares | Mean Square | $F_c$ |
|---|---|---|---|---|
| Regression | k | $\Sigma(\hat{Y}-\overline{Y})^2=\Sigma\hat{y}^2$ = RSS | A = RSS/k | |
| Residual (error) | n-k-1 | $\Sigma(Y-\hat{Y})^2=\Sigma e^2$ = SSE = TSS-RSS | B = SSE/n-k-1 | A/B |
| Total | n-1 | $\Sigma(Y-\overline{Y})^2=\Sigma y^2$ = TSS $=\Sigma Y^2-\frac{(\Sigma Y)^2}{n}$ | | |

n = จำนวนข้อมูล

k = จำนวนตัวแปรอิสระ

เปรียบเทียบค่า $F_c$ ที่ได้จากการคำนวณเทียบกับ $F_t$ ที่ได้จากตาราง F-distribution ที่ d.f = (k,n-k-1) ณ ระดับความเชื่อมั่นต่าง ๆ เช่น $\alpha$ = 0.05, 0.01 ทำให้สามารถที่จะสรุปผลของการวิเคราะห์ได้ว่า $X_1$ , $X_2$ , $X_3$ มีอิทธิพลต่อ Y อย่างมีนัยสำคัญหรือไม่

การคำนวณหาค่า t - value เพื่อทดสอบค่า $\beta_i$ โดยการตั้งสมมติฐาน

$$H_o : \beta_i = 0, \quad i = 1,2.....k$$

$$t_c = \frac{b_i - \beta_i}{S_{b_i}} \quad , d.f = n-k-1$$

$$= \frac{b_i}{S_{b_i}}$$

$b_i$ เป็นค่า Regression Coefficient ที่ต้องการทดสอบ

$t_c$ เป็นค่าที่คำนวณได้จากสูตร

$S_{b_i}$ เป็นค่า Standard Error ของ $b_i$

$$S_{b_i} = \sqrt{C_{ii}\ S^2_{Y.1\ .....k}}$$

$C_{ii}$ เป็นค่าของ diagonal จาก $(X' X)^{-1}$

$S^2_{Y.1.....k}$ เป็นค่า Standard Error ของการประมาณค่า $\hat{Y}$ ซึ่งเป็นค่าของ Mean Square of deviation ในตาราง ANOVA ของ Multiple Linear Regression

$$S_{Y.1.....k} = \sqrt{\frac{\Sigma\ (Y - \hat{Y})^2}{n - k - 1}}$$

จากนั้นนำค่า $t_c$ ที่คำนวณได้เปรียบเทียบกับค่า t จากตาราง t-distribution ที่ d.f = n - k - 1 ณ ระดับความเชื่อมั่นต่าง ๆ แล้วจึงจะสามารถสรุปผลได้ว่าจะยอมรับหรือปฏิเสธสมมติฐาน $H_o : \beta_i = 0$

๓.๓.๒ สัมประสิทธิ์แห่งสหสัมพันธ์ (Coefficient of Correlation)

สัมประสิทธิ์แห่งสหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) เป็นค่าสัมพัทธ์ (Relative Value) แสดงความสัมพันธ์ระหว่างตัวแปร (Variables) ทั้งสองว่ามีอยู่มากน้อยเพียงใด ตัวแปรทั้งสองปกติจะใช้แทนด้วย X กับ Y และค่าของสัมประสิทธิ์แห่งสหสัมพันธ์จะแทนด้วย $r_{XY}$ หรือ $R_{XY}$

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

$$S_X = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{(n - 1)}}\ ,\ S_Y = \sqrt{\frac{\Sigma(Y_i - \overline{Y})^2}{(n - 1)}}$$

$$S_{XY} = \frac{\Sigma\ (X_i - \overline{X})\ (Y_i - \overline{Y})}{(n - 1)}$$

$S_X$ , $S_Y$ เป็นค่าเบี่ยงเบนมาตรฐาน (Standard deviation ) ของ X และ Y

$S_{XY}$ เป็น Covariance ของ X กับ Y

$$\therefore \quad r_{XY} = \frac{\sum ( X_i - \overline{X} ) ( Y_i - \overline{Y} )}{\sqrt{\sum ( X_i - \overline{X} )^2 \sum ( Y_i - \overline{Y})^2}}$$

การพิจารณาค่าของ r มี ๓ กรณี

๑. r มีค่าเป็นบวก หมายความว่าค่าของตัวแปรทั้งคู่จะเปลี่ยนไปในลักษณะเดียวกันคือ เมื่อ X มีค่าเพิ่มขึ้น Y ก็จะมีค่าเพิ่มขึ้นด้วย

๒. r มีค่าเป็นลบ หมายความว่าค่าของตัวแปรทั้งคู่จะเปลี่ยนไปในลักษณะตรงกันข้าม กล่าวคือ X มีค่าเพิ่มขึ้น ค่าของ Y ก็จะมีค่าลดลง

๓. r มีค่าเข้าใกล้กับ ๐ หมายความว่า X กับ Y ไม่มีความสัมพันธ์ต่อกัน ค่าของ r จะมีค่าอยู่ระหว่าง - 1 กับ + 1 ; $-1 \leq r \leq 1$

และค่า $|r|$ เป็นค่าสมบูรณ์ของ r ถ้ามีค่าเข้าใกล้ 1 แสดงว่า X กับ Y มีความสัมพันธ์ต่อกันมาก ถ้ามีค่าเข้าใกล้ ๐ แสดงว่า X กับ Y เกือบไม่มีความสัมพันธ์กัน

ดัชนีกำหนดหรือสัมประสิทธิ์แห่งการตกลงใจ (Coefficient of determination)

คือ $$r_{XY}^2 = \frac{\left\{\sum (X_i - \overline{X} ) (Y_i - \overline{Y} )\right\}^2}{\sum ( X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}$$

$r^2$ เรียกว่าดัชนีกำหนด ส่วนที่เหลือ $1 - r^2$ เรียกว่าดัชนีไม่กำหนด

(Coefficient of non determination )

เมื่อให้ $x = (X - \bar{X})$ , $y = (Y - \bar{Y})$

ดังนั้น $$r^2 = \frac{(\sum xy)^2}{\sum x^2 \sum y^2}$$

$$r^2 = \frac{(\sum xy)^2 / \sum x^2}{\sum y^2}$$

หรือ $$r^2 = \frac{(\sum xy)^2 / \sum y^2}{\sum x^2}$$

<u>ค่าสัมประสิทธิ์สหสัมพันธ์เชิงส่วน</u> (Partial Correlation Coefficient)

ในกรณีที่มีตัวแปรอิสระหลายตัวอยู่ในสมการถดถอย นอกจากจะหาความสัมพันธ์กันเป็นคู่ ๆ คือหาค่าสหสัมพันธ์อย่างง่ายระหว่างตัวแปร $X_i$ กับ $X_j$ มาแล้ว ถ้าตัวแปรต่าง ๆ ทุกตัวฝันแปรร่วมกันแบบการแจกแจงปกติ (Normal Distribution) เราก็อาจจะหาความสัมพันธ์ในลักษณะสหสัมพันธ์เชิงส่วนของตัวแปรได้ เช่น

$$Y = b_o + b_1 X_1 + b_2 X_2 + b_3 X_3$$

นอกจากจะได้ค่า $r_{YX_1}$, $r_{YX_2}$ , $r_{YX_3}$ แล้วเรายังอาจหาค่าสหสัมพันธ์เชิงส่วนได้ เช่น

$r_{YX_1 \cdot X_2}$ หรือ $r_{Y1 \cdot 2}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับ $X_1$ โดยถือว่า $X_2$ คงที่

$r_{YX_2 \cdot X_1}$ หรือ $r_{Y2 \cdot 1}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับ $X_2$ โดยถือว่า $X_1$ คงที่

$r_{YX_3 \cdot X_1}$ หรือ $r_{Y3 \cdot 1}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับ $X_3$ โดยถือว่า $X_1$ คงที่

$r_{YX_2 \cdot X_3}$ หรือ $r_{Y2 \cdot 3}$ คือค่าสหสัมพันธ์เชิงส่วนระหว่างตัวแปร Y กับ $X_2$ โดยถือว่า $X_3$ คงที่

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\left[(1 - r_{13}^2)\ (1 - r_{23}^2)\right]^{1/2}}$$

เรียก $r_{12.3}$ ว่าเป็น Partial Correlation ค่าของ r และเครื่องหมายมีความหมายเช่นเดียวกับกรณีค่าสหสัมพันธ์อย่างง่าย

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\left[(1 - r_{Y2}^2)\ (1 - r_{12}^2)\right]^{1/2}}$$

$$r_{Y3.12} = \frac{r_{Y3.1} - r_{Y2.1}\, r_{32.1}}{\left[(1 - r_{Y2.1}^2)\ (1 - r_{32.1}^2)\right]^{1/2}}$$

ค่าสัมประสิทธิ์สหสัมพันธ์เชิงซ้อน (Multiple Correlation Coefficient)

โดยทั่วไปเป็นค่าสหสัมพันธ์ระหว่าง Y และ $X_1, X_2$ ..... Xk หรือเป็นค่าสหสัมพันธ์อย่างง่ายระหว่าง Y กับ $\hat{Y}$ ใช้ R

$$R = r_{Y\hat{Y}} = \frac{\mathrm{Cov}\,(Y,\hat{Y})}{\left[V\,(Y)\ V\,(Y)\right]^{1/2}} \quad ; \ Y = \hat{Y} + e$$

$$R^2 = r_{Y\hat{Y}}^2$$

เศษของ $\mathrm{Cov}\,(Y,\hat{Y}) = \Sigma\, y\ \hat{y}$

$$= \Sigma\, \hat{y}^2$$

ดังนั้น $r_{y\hat{y}} = \frac{= \Sigma\, \hat{y}^2}{\left[\Sigma\, y^2 \Sigma\, \hat{y}^2\right]^{1/2}}$

$$R^2 = \frac{\Sigma\, \hat{y}^2}{\Sigma\, y^2}$$

$$1 - R^2 = 1 - \frac{\Sigma \hat{y}^2}{\Sigma y^2}$$

การตรวจสอบนัยสำคัญของค่าสหสัมพันธ์เชิงซ้อนโดยการตั้งสมมติฐาน

$$Hc : B_1 = B_2 = \cdots\cdots = B_k = 0$$

แล้วใช้ F - test ซึ่งมีค่า

$$F = \frac{(n - k - 1) R^2}{k (1 - R^2)} \qquad \text{ใช้} \quad d.f = (k, n - k - 1$$

สมมติว่า Y เกี่ยวข้องกับตัวแปรอิสระที่สำคัญ ๓ ตัว เช่น $X_2$ , $X_4$ และ $X_5$

ค่าของ Multiple Standard Error of Estimate

$$S_{Y.245} = \sqrt{\frac{\Sigma (Y - \hat{Y})^2}{(n-k-1)}} = \sqrt{\frac{\Sigma (Y - \hat{Y})^2}{10}}$$

n = จำนวนข้อมูล = ๑๔

k = จำนวนตัวแปรอิสระ = ๓

๓.๓.๓ การสร้างสมการถดถอยเชิงซ้อนโดยคอมพิวเตอร์

ในการหาโมเดลที่เหมาะสมสำหรับปริมาณการผลิตน้ำตาลทรายและปริมาณการบริโภคน้ำตาลทรายภายในประเทศ จะอาศัยวิธี Stepwise Multiple Regression โดย เลือกตัวแปรอิสระ ที่มีอิทธิพลต่อตัวแปรตามมากที่สุด และรอง ๆ ลงมาเข้ามาในสมการ เพื่อให้ได้สมการที่แสดงค่าประมาณของตัวแปรตามได้ ใกล้เคียงความเป็นจริงมากที่สุด

วิธีการเลือกตัวแปรอิสระเข้าในสมการหรือโมเดลกระทำดังนี้ :-

๑. พิจารณาเลือกก่อน $X_i$ เข้าในสมการทีละตัวตามลำดับความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ $X_i$ กล่าวคือจะต้องเป็น $X_i$ (ตัวแปรอิสระ) ตัวที่ให้ค่าดัชนีกำหนด ($r^2$) สูงสุดเข้าเป็นตัวแรก

เช่น สมมติให้ $X_1$ เป็นตัวที่ให้ค่าดัชนีกำหนดสูงสุดคือ $V_1 = r^2_{YX_1}$ มีค่าสูงสุดดังนั้น จะได้สมการในขั้นแรกเป็น $\hat{Y} = a^I + b^I_1 X_1$

๒. ตัวแปรอิสระตัวที่สองที่เลือกเข้าไปในโมเดล คือตัวแปรอิสระที่ให้ค่า ผลคูณระหว่างค่าดัชนีกำหนด ของตัวมันเองกับค่าดัชนีกำหนดที่เหลือจากตัวแปรอิสระตัวแรก สูงสุด สมมติว่าเลือกได้เป็น $X_2$ นั่นคือ $X_2$ จะให้ค่า

$V_2 = (1-r^2_{YX_1})\, r^2_{YX_2 \cdot X_1}$ สูงสุด

ดังนั้นจะได้สมการใหม่ในขั้นที่ ๒ เป็น $\hat{Y} = a^{II} + b^{II}_1 X_1 + b^{II}_2 X_2$

๓. ตัวแปรอิสระตัวต่อไปสมมติได้ $X_3$ นั่นคือ $X_3$ จะให้ค่า

$V_3 = (1-r^2_{YX_1})\,(1-r^2_{YX_2 \cdot X_1})\, r^2_{YX_3 \cdot X_1 X_2}$ สูงสุด

ได้สมการใหม่เป็น $\hat{Y} = a^{III} + b^{III}_1 X_1 + b^{III}_2 X_2 + b^{III}_3 X_3$

ทำเช่นนี้ต่อไปจนครบ $k$ ตัวจะได้

$$V_k = (1-r^2_{YX_1})(1-r^2_{YX_2 \cdot X_1}) \cdots (1-r^2_{YX_{k-1} \cdot X_1 X_2 \cdots X_{k-2}})\, r^2_{YX_k \cdot X_1 X_2 \cdots X_{k-1}}$$

โดยมีสมการเป็น

$$\hat{Y} = a^{(k)} + b^{(k)}_1 X_1 + b^{(k)}_2 X_2 + \cdots + b^{(k)}_k X_k$$

เนื่องจากตัวแปรอิสระทุกตัวไม่ได้มีอิทธิพลอย่างมีนัยสำคัญต่อโมเดลทุกตัว. ดังนั้น โมเดลในขั้นสุดท้าย ซึ่งมีตัวแปรอิสระ $k$ ตัวจึงอาจจะไม่ใช่โมเดลที่มีนัยสำคัญเสมอไป วิธีการที่จะเลือกว่าสมการในขั้นตอนใดควรจะยุติลงได้ จะยังคงใช้หลักในการพิจารณาเช่นเดียวกับกรณีการเลือกหาค่าแนวโน้มตามลำดับเวลาก่อนหน้านี้คือ :-

๑. พิจารณาว่าสมการขั้นใดที่ให้ค่า Standard error ต่ำสุดและ ค่าดัชนีกำหนดสูงสุด (เมื่อเพิ่มตัวแปรอิสระเข้าไปอีกค่า S.E. จะเพิ่มขึ้น)

๒. พิจารณาจากการทดสอบนัยสำคัญของสมการโดยพิจารณา F-test จาก ANOVA

ANOVA

๓. พิจารณาจากการทดสอบนัยสำคัญของ Partial Regression Coefficient โดย t-test