บทที่ 4 .

การวิเคราะห์การถดถอย

4.1 การวิเคราะห์การถดถอยกับงานวิจัย

เนื่องจากงานวิจัยนี้มีจุดประสงค์อย่างหนึ่ง คือใช้การถดถอยพหุคูณ (Multiple regression) วิเคราะห์ข้อมูลที่เก็บจากสถานที่ทำงานทั้ง 4 แห่ง เพื่อหาสมการโหวตเฉลี่ยของคนไทย (ในที่นี้เป็นตัวแปรตาม) ว่ามีตัวแปรอิสระอะไรบ้างที่มีนัยสำคัญ(Significance)

ตัวแปรตามและตัวแปรอิสระที่ใช้ในการวิเคราะห์ในงานวิจัยมีดังนี้

ตัวแปรตามได้แก่ PMV หรือ ดัชนีทำนายการโหวตเฉลี่ย

ตัวแปรอิสระมี 6 ตัว ได้แก่

1 ลักษณะงาน
2 ความต้านทานความร้อนของเสื้อผ้า หรือชุดแต่งกาย
3 อุณหภูมิอากาศในห้องทดสอบ
4 อุณหภูมิการแผ่ความร้อนเฉลี่ย
5 ความเร็วอากาศสัมพัทธ์
6 ความดันไอน้ำในอากาศ

ดังนั้นจะขอกล่าวถึงทฤษฎีการถดถอยอย่างง่ายและพหุคูณ ดังหัวข้อต่อไปนี้

4.2 การถดถอยและสหสัมพันธ์ (Regression and correlation)

หมายถึง วิธีการอย่างหนึ่งที่ใช้ในการศึกษาตัวแปรหนึ่งเมื่อมีความสัมพันธ์ หรือได้รับอิทธิพลจากตัวแปรอื่นๆ ตัวแปรที่เกี่ยวข้องกับตัวแปรที่ต้องการศึกษาอาจมีตัวเดียว การวิเคราะห์ลักษณะนี้เรียกว่า การวิเคราะห์การถดถอยและสหสัมพันธ์อย่างง่าย (Simple regression and correlation analysis) แต่ในบางกรณีตัวแปรที่เกี่ยวข้องกับตัวแปรที่

ต้องการศึกษาอาจมีตั้งแต่ 2 ตัวขึ้นไป การวิเคราะห์การถดถอย ลักษณะนี้ เรียกว่า การถดถอยพหุคูณ

ไม่ว่าจะเป็นการวิเคราะห์การถดถอยอย่างง่าย หรือ การถดถอยพหุคูณ วิธีการที่ใช้คล้ายคลึงกัน เพียงแต่การวิเคราะห์การถดถอยพหุคูณ วิธีการจะยุ่งยากขึ้นเนื่องจากตัวแปรที่เกี่ยวข้องมีมากกว่า 1 ตัว ยิ่งตัวแปรที่เกี่ยวข้องมีมากเท่าใด ความยุ่งยากก็จะเพิ่มมากขึ้นเท่านั้น

4.3 ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

การวิเคราะห์การถดถอยอย่างง่ายมีข้อตกลงเบื้องต้นดังนี้

4.3.1 ตัวแปรอิสระ X และตัวแปรตาม Y จะต้องมีความสัมพันธ์ในแบบเส้นตรง

$$Y = a + bX + e$$

เมื่อ e คือค่าคลาดเคลื่อนสุ่มที่เกิดขึ้นในค่าที่สังเกตุได้ของตัวแปร Y อันเนื่องมาจากตัวอย่างที่สุ่มมา

4.3.2 ตัวแปรตามต้องเป็นตัวแปรสุ่มชนิดต่อเนื่อง ในขณะที่ตัวแปรอิสระเป็นเซ็ทของค่าต่าง ๆ ที่เลือกกำหนดได้

4.3.3 ความแปรปรวนของ Y สำหรับแต่ละค่าของ X ที่กำหนดจะต้องมีค่าเท่ากัน หรือเรียกว่า Homoscedasticity สามารถพิจารณาอย่างคร่าว ๆ ว่าข้อมูลที่ได้มาสอดคล้องกับคุณสมบัติข้อนี้หรือไม่อาจทำได้โดยการพิจารณาจากแผนภาพการกระจายของข้อมูลจากตัวอย่าง ถ้าการกระจัดกระจายของจุดไปจากเส้นแนวโน้มมีพอ ๆ กัน สำหรับค่าต่าง ๆ ของ X ที่กำหนด แสดงว่าข้อมูลที่ได้มาสอดคล้องกับคุณสมบัติข้อนี้

4.3.4 ค่าที่สังเกตุได้แต่ละค่าต้องไม่มีความสัมพันธ์กัน ถ้าเป็นตัวอย่างที่สุ่มมาก็จะสอดคล้องกับคุณสมบัติข้อนี้

4.4 การถดถอยเชิงเส้นอย่างง่าย(Simple linear regression)

มีสมการพื้นฐานคือ

$$Y' = a + bX \qquad (4.1)$$

เมื่อ Y' คือ ค่าทำนายตัวแปรตาม Y

X คือ ค่าตัวแปรอิสระ

a คือ ค่าคงที่ Intercept

b คือ สัมประสิทธิ์การถดถอย

โดยที่ ค่า a หาได้จากสูตร

$$a = \bar{Y} - b\bar{X} \qquad (4.2)$$

และ b หาได้จากสูตร

$$b = \Sigma xy / \Sigma x^2 \qquad (4.3)$$

เมื่อ x,y คือ ค่าเบี่ยงเบนจากค่าเฉลี่ยของ X,Y หรือ $x = X - X$ $y = Y - Y$ และ $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N$, $\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/N$, $\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/N$ และ N คือ จำนวนคู่ข้อมูล

ส่วนวิธีการลากเส้นถดถอย มีกฎอยู่ว่า ต้องลากผ่านค่าคงที่ Intercept และจุด X,Y

4.5 ความหมายและการหาค่าสัมประสิทธิ์สหสัมพันธ์ (r)

ในการวิเคราะห์การถดถอย จะมีสัมประสิทธิ์ตัวหนึ่งที่เป็นดัชนี แสดงขนาดและทิศทางของความสัมพันธ์ระหว่างตัวแปร ดัชนีตัวนั้นได้แก่ สัมประสิทธิ์สหสัมพันธ์ (r) ซึ่งค่า r มีค่าอยู่ระหว่าง $-1 < r < +1$ ถ้า $r = -1.0$ เมื่อนำมาพล็อตกราฟ ค่า X,Y ทุกค่าจะอยู่บนเส้นถดถอย แต่ถ้า r ยิ่งเข้าใกล้ 0 จุด X,Y จะยิ่งกระจายห่างจากเส้นถดถอยมากขึ้น เครื่องหมาย + หรือ - ติดอยู่ แสดงว่า slope ของเส้นถดถอยเป็น + หรือ - ส่วนการหาค่าสัมประสิทธิ์ตัวนี้มีหลายวิธี แต่ละวิธีจะให้ค่าเท่ากันเช่น

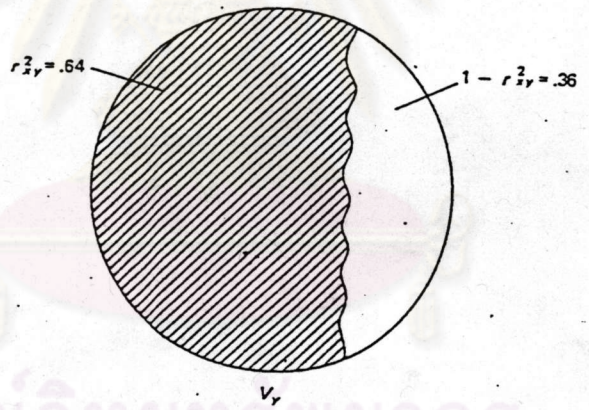
$$r_{xy} = \Sigma xy / (\Sigma x^2 \Sigma y^2)^{0.5} \qquad (4.4)$$

$$r_{xy} = \Sigma xy / Ns_x s_y \qquad (4.5)$$

เมื่อ S_x คือ ค่าเบี่ยงเบนมาตราฐานของ X

4.6 ความหมายและการหาค่าสัมประสิทธิ์การตัดสินใจ (r^2)

เป็นสัมประสิทธิ์ที่จะบอกให้ทราบถึงความแปรปรวนของ Y ที่สามารถอธิบายด้วยตัวแปร X ได้มากน้อยเท่าไร เช่น



รูปที่ 4.1 สัดส่วนความแปรปรวนของ Y

จากรูปที่ 4.1 พื้นที่แรงเงาคือสัดส่วนของความแปรปรวน Y ที่สามารถอธิบายได้ด้วยตัวแปร X = 0.64 ไม่สามารถอธิบายด้วยตัวแปร X = 0.36 เมื่อพื้นที่วงกลม (V_y) คือ ความแปรปรวนทั้งหมดของ Y = 1

4.7 การทดสอบนัยสำคัญของการถดถอยอย่างง่าย

เนื่องจากเส้นถดถอยสำหรับชุดข้อมูลแต่ละชุดเป็นเพียงเส้นหนึ่งที่ถูกเลือกขึ้นมา ดังนั้นความคลาดเคลื่อนที่เกิดขึ้นจึงเป็นแบบสุ่ม จึงจำเป็นต้องมีการทดสอบนัยสำคัญของสัมประสิทธิ์ถดถอย

โดยการตั้งสมมุติฐาน เมื่อนำมาใช้กับการวิเคราะห์การถดถอยซึ่งมีรูปสมการดังนี้

$$SS_t = SS_{reg} + SS_{res} \qquad (4.6)$$

เมื่อ SS_t คือ ผลรวมทั้งหมดกำลังสองของ Y

$= \Sigma y_t = \Sigma Y^2 - (\Sigma Y)^2/N$

SS_{reg} คือ ผลรวมกำลังสองของ Y เนื่องจากการถดถอย

$= \Sigma y'^2 = \Sigma Y'^2 - (\Sigma Y')^2/N$

SS_{res} คือ ผลรวมกำลังสองของ Residuals

$= \Sigma d^2 = \Sigma(Y-Y')^2$

และการทดสอบหานัยสำคัญจะใช้ F-ratio ดังสูตร

$$F = (SS_{reg}/df_1)/(SS_{res}/df_2) \qquad (4.7)$$

เมื่อ df_1 คือ องศาอิสระ(degree of freedom)ของการถดถอย(=K)

df_2 คือ องศาอิสระของresiduals(=N-K-1)

K คือ จำนวนตัวแปรอิสระ

เมื่อหาค่า F ได้แล้วจะเปิดตาราง F- distributionในภาคผนวกเพื่อตรวจสอบดูว่ามีนัยสำคัญหรือไม่

4.8 การวิเคราะห์การถดถอยพหุคูณ(Multiple regression analysis)

วิธีวิเคราะห์ที่จะกล่าวต่อไปนี้เป็น General method ของการวิเคราะห์ฯ ที่มีตัวแปรอิสระกี่ตัวก็ได้ สามารถใช้ได้กับการถดถอยทุกรูปแบบไม่ว่าตัวแปรจะเป็นแบบ Cotineous , Category หรือ Codes

4.9 การวิเคราะห์สมการที่มีตัวแปรอิสระ 3 ตัวขึ้นไป

ในกรณีที่มีตัวแปรอิสระ 3 ตัว จำเป็นต้องใช้ Matrix หา สปส.การถดถอย (แต่ถ้ามีตัวแปรอิสระ 2 ตัว ไม่จำเป็นต้องใช้ matrix ก็สามารถ Solve ปัญหาได้ ในที่นี้จะไม่ขอกล่าวถึง) รูปแบบของสมการการถดถอยพหุคูณทั่วไปเป็นดังนี้

$$Y' = a + b_1X_1 + b_2x_2 + \ldots + b_kx_k \qquad (4.8)$$

ค่า $a, b_{1\ldots k}$ เป็นค่าคงที่ที่ต้องหาเพื่อให้สมการมี residuals น้อยที่สุด

4.10 ข้อสมมุติฐานในการวิเคราะห์การถดถอยพหุคูณ

ในการวิเคราะห์การถดถอยพหุคูณเชิงเส้นจะมีสมมุติฐานดังนี้

4.10.1 ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในแบบเส้นตรง สมการที่มีตัวแปรอิสระ k ตัว จะอยู่ในรูป

$$Y = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots b_kX_k + e \qquad (4.9)$$

เมื่อ e คือ ความคลาดเคลื่อนสุ่มซึ่งเกิดจากการสุ่มตัวอย่าง

4.10.2 ตัวแปรตามเป็นตัวแปรสุ่มต่อเนื่อง ในขณะที่ตัวแปรอิสระเป็นตัวแปรควบคุมและเป็นเซ็ทของค่าซึ่งไม่ใช่ค่าสุ่ม ดังนั้น ค่า e ในรูปแบบเชิงเส้นจึงเป็นความคลาดเคลื่อนของตัวแปรตามอันเกิดจากการสุ่มตัวอย่าง การวิเคราะห์การถดถอยพหุคูณสามารถนำมาใช้ได้แม้ว่าตัวแปรอิสระจะเป็นตัวแปรสุ่มเช่นเดียวกัน

4.10.3 ความแปรปรวนของตัวตามเมื่อกำหนดค่าต่าง ๆ ของตัวแปรอิสระจะมีค่าเท่ากันหรือเรียกว่ามีคุณสมบัติของ Homoscedasticity

4.10.4 ค่าของตัวแปรสุ่มตัวที่อยู่ติดกันไม่สัมพันธ์กัน ถ้าค่าของตัวแปรสุ่มไม่สอดคล้องกับคุณสมบัติข้อนี้ เรียกว่า ข้อมูลมี Autocorrelation กัน

4.10.5 ตัวแปรอิสระแต่ละตัวจะต้องไม่มีความสัมพันธ์กัน

ถ้าจะกล่าวกันโดยแท้จริงแล้ว ข้อสมมุติฐานทั้งห้า เป็นสิ่งที่จะสมบูรณ์ได้ยาก หากข้อสมมุติฐานดังกล่าวเป็นสิ่งใกล้เคียงกับความจริง แบบแผนของความสัมพันธ์ก็จะเป็นสิ่งที่มีประโยชน์มาก

4.11 Solution ทั่วไปของสมการการถดถอย

จากสมการ (4.8) เราต้องมีสมการ k สมการในการหาค่า b

จาก Calculus จะได้สมการปกติ (Normal equation) สำหรับตัวแปรอิสระ 3 ตัว ดังนี้

$$\begin{aligned} \beta_1 + r_{12}\beta_2 + r_{13}\beta_3 &= r_{y1} \\ r_{21}\beta_1 + \beta_2 + r_{23}\beta_3 &= r_{y2} \\ r_{31}\beta_1 + r_{32}\beta_2 + \beta_3 &= r_{y3} \end{aligned} \qquad (4.10)$$

หมายเหตุ 1. $r_{12} = r_{21}$, $r_{13} = r_{31}$ ฯลฯ และ r_{11} , r_{22} , r_{33} จะไม่เขียนในสมการเพราะมีค่าเท่ากับ 1

2. β คือ สัมประสิทธิ์การถดถอย population

จาก (4.10) สามารถเขียนเป็น matrix ได้ดังนี้

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \end{bmatrix} \qquad (4.11)$$

หรือ $[\, R_{ij} \,][\, \beta_j \,] = [\, R_{yj} \,]$

เป็นที่ทราบแล้วว่า matrices สามารถ บวก, ลบ, คูณ ได้ แต่ไม่สามารถนำมาหารได้แต่จะใช้ inverse matrix แทน

จาก (4.11) จะได้

$$[\beta_j] = [R_{ij}]^{-1}\,[R_{yi}] \qquad (4.12)$$

เมื่อ $[R_{ij}]^{-1}$ คือ inverse ของ $[R_{ij}]$

จาก(4.12) จะหาค่า b จากสูตร

$$b_j = \beta_j(s_y/s_j) \qquad (4.13)$$

เมื่อ b คือ regression weight ที่ j=1,2,3 ...
s_y คือ ค่าเบี่ยงเบนมาตราฐานของตัวแปรตาม Y
s_j คือ ค่าเบี่ยงเบนมาตราฐานของตัวแปรอิสระ

4.12 ความหมายและการหาค่าสัมประสิทธิ์การตัดสินใจพหุคูณ (R^2)

คือสัมประสิทธิ์ที่บอกถึงสัดส่วนของความแปรปรวนในตัวแปรตาม Y ที่อธิบายได้ด้วยตัวแปรอิสระต่าง ๆ ในเส้นถดถอย หาได้จากสูตร

$$R^2 = SS_{reg} / SS_t \qquad (4.14)$$

เมื่อ SS_{reg} หาได้จาก $\Sigma y'^2$
SS_{res} หาได้จาก Σd^2

4.13 ความหมายและการหาค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณ (R)

คือ สัมประสิทธิ์สหสัมพันธ์เชิงพหุคูณที่อยู่ในรูปสัมบูรณ์หรือไม่มีเครื่องหมายกำกับ เพราะอาจเป็นไปได้ที่ตัวแปรตัวหนึ่งมีความสัมพันธ์ในทางตรงกับตัวแปรตามในขณะที่ตัวแปรอื่น ๆ มีความสัมพันธ์แบบผกผันกับตัวแปรตาม ดังนั้นความสัมพันธ์ระหว่างตัวแปรทั้งหมดจึงไม่สามารถบอกได้ชัดเจนว่าเป็นไปในทางบวกหรือลบ จะหาได้จากสูตร

$$R = (R^2)^{0.5} \qquad (4.15)$$

เช่น $R^2 = 0.66$
จะได้ $R = 0.81$

4.14 การทดสอบนัยสำคัญของการถดถอยพหุคูณ

หลังจากตั้งสมมุติฐานแล้วจะใช้การทดสอบ F-ratio ตรวจสอบการมีนัยสำคัญ โดยใช้สูตร

$$F = \frac{R^2/K}{(1-R^2)/(N-K-1)} \qquad (4.16)$$

เมื่อ K = จำนวนตัวแปรอิสระ

เมื่อหาค่า F ได้แล้วจะเปิดตาราง F-distribution ในภาคผนวกเพื่อตรวจสอบว่ามีนัยสำคัญหรือไม่

4.15 การทดสอบนัยสำคัญทางสถิติของตัวแปรที่เพิ่มเข้ามาในสมการถดถอย

จากหัวข้อ 4.14 เรายังไม่ทราบว่า X ตัวไหนมีอิทธิพลต่อ Y มากหรือน้อย (เนื่องจากมี X หลายตัว) ดังนั้น จึงต้องมีการทดสอบฯ ว่า X ใดมีนัยสำคัญบ้าง โดยลดตัวแปรอิสระลงทีละตัว แล้วตรวจสอบนัยสำคัญของตัวแปรนั้นว่ามีหรือไม่ ถ้าไม่มีแสดงว่าเราสามารถตัดตัวแปรนั้นออกจากสมการได้

หมายเหตุ เนื่องจากการวิเคราะห์การถดถอย เป็นวิธีการทางสถิติที่นำมาใช้ในงานวิจัยนี้ เพื่อให้มีความเข้าใจได้ง่ายขึ้น โปรดดูตัวอย่างการวิเคราะห์ฯ ในภาคผนวก ข.1