บทที่ 4
สหสัมพันธ์

4.1 ความนำ

สหสัมพันธ์คือดัชนีวัดความผันแปรร่วมกันระหว่างตัวแปรสุ่ม ในทางปฏิบัติเราอภิปรายความหมายของสหสัมพันธ์ในลักษณะของดัชนีที่วัดปริมาณของความข้องเกี่ยวกันของตัวแปร โดยถือว่าถ้าสหสัมพันธ์มีค่าเป็นตัวเลขในปริมาณสูงถือว่าตัวแปรมีความข้องเกี่ยวกันมาก และความข้องเกี่ยวจะค่อย ๆ ลดลงถ้าค่าตัวเลขมีปริมาณ (magnitude) ลดลง ส่วนเครื่องหมายบวกหรือลบเป็นดัชนีที่ชี้ให้เห็นว่าตัวแปรที่สนใจนั้นมีการผันแปรไปในทิศทางเดียวกันหรือว่าสวนทางกัน ทั้งนี้มิได้หมายความว่าใครเป็นเหตุใครเป็นผล หากเพียงแต่แสดงให้ทราบว่าสัมพันธ์กันมากน้อยเพียงใดในทิศทางใดเท่านั้น เช่น

$r_{12} = r_{x_1x_2} = .85$ แสดงว่าตัวแปรสุ่ม X_1 และ X_2 มีความสัมพันธ์กันค่อนข้างสูง โดยที่ X_1 และ X_2 จะผันแปรไปในทิศทางเดียวกันแต่ทั้งนี้ไม่ได้แสดงว่า X_1 ผันแปรไปเพราะสาเหตุจาก X_2 หรือ X_2 ผันแปรไปเพราะ X_1

$R_{12\cdot3} = r_{x_1x_2\cdot x_3} = -0.77$ แสดงว่าเมื่อเราควบคุมให้ X_3 มีอิทธิพลคงที่ (เช่นกำหนดให้ $X_3 = c$)[1] X_1 และ X_2 จะมีความเกี่ยวข้องกันในอัตราค่อนข้างสูงแต่มีลักษณะของการผันแปรในทิศทางที่สวนทางกัน

$R_{x_1\,x_2\cdot x_3}$ เรียกว่า Partial Correlation ในทางทฤษฎีเราเสนอในรูปทั่วไปดังนี้คือ $R_{ij\cdot q+1,\ q+2,\ldots,\ k}$ โดย Partial Correlation ถูกนำเข้ามาใช้เพื่อช่วยให้มองเห็นภาพของความสัมพันธ์ระหว่างตัวแปรที่เสนอในรูป Simple Correlation คือ r_{ij} กระจ่างชัดขึ้น ตัวอย่างเช่นเราต้องการศึกษาสหสัมพันธ์ระหว่างรายได้ของครูกับยอดจำหน่ายสุราแม่โขง จากการวิเคราะห์ข้อมูลพบว่า $r_{x_1x_2} = .98$ (X_1=รายได้ของครู, X_2 = ยอดขายสุราแม่โขง) เมื่อพิจารณาจากค่าสหสัมพันธ์จะพบ

---

[1] ในทางทฤษฎีค่า $R_{x_1x_2\cdot x_3}$ คำนวณจาก $f(X_1, X_2 \mid X_3)$ และในทางปฏิบัติเราหมายถึงการควบคุมอิทธิพลของ X_3 ไว้โดยกำหนดให้ $X_3 = c$ ขอให้เข้าใจว่า $f(X_1, X_2 \mid X_3) = \dfrac{f(X_1, X_2, X_3)}{f(X_3)}$ ไม่หมายความว่าขจัดอิทธิพลของ X_3 ทิ้ง

ว่ารายได้ของครูและยอดขายสุราแม่โขงมีอัตราความสัมพันธ์กันค่อนข้างสูง [1] แต่โดยข้อเท็จจริงแล้วรายได้ของครูและยอดขายสุราแม่โขงไม่น่าจะสัมพันธ์กันได้ เราจึงวิเคราะห์ว่าการที่ตัวแปรคู่นี้สัมพันธ์กันได้จะต้องมีสาเหตุอื่นแฝงเร้นอยู่ซึ่งจะเป็นตัวการที่แอบส่งอิทธิพลให้เกิดค่าสหสัมพันธ์ที่ลวงตาเกิดขึ้นดังกล่าวและถ้าเราควบคุมอิทธิพลของตัวแปรที่แฝงเร้นอยู่นั้นเสียได้ ภาพความสัมพันธ์ที่แท้จริงระหว่างรายได้ของครูและยอดขายสุราแม่โขงก็จะปรากฏขึ้น สมมุติว่าตัวแปรที่แฝงเร้นคือ X_3 = รายได้ต่อบุคคล X_4 = รสนิยมของผู้บริโภค เมื่อเราควบคุม X_3 และ X_4 ให้คงที่อยู่ ณ ค่าคงที่ใด ๆ เราก็ย่อมสามารถหาค่าสหสัมพันธ์ที่แท้จริงของ X_1 และ X_2 ได้ โดยนัยดังกล่าวนี้ Partial Correlation จึงเข้ามามีบทบาทในงานวิเคราะห์ข้อมูลเพิ่มเติมไปจากการมีเฉพาะ Simple Correlation ตามตัวอย่างนี้ Partial Correlation ก็คือ $R_{x_1x_2 \cdot x_3x_4}$

ทั้ง Simple Correlation และ Partial Correlation มีความสำคัญเป็นอย่างมากในการสร้าง Best Fitted Model โดยเฉพาะในกรณีของ Stepwise Regression นอกเหนือไปจากนี้เฉพาะ Simple Correlation ยังมีความสำคัญเป็นอย่างยิ่งในการวิเคราะห์ปัญหา Multicolinearity และยังสามารถใช้ในการประมาณค่า β ได้อีกด้วย ความจำเป็นของการศึกษาเรื่องราวของสหสัมพันธ์จึงยังคงต้องมีอยู่แม้ว่าสัมประสิทธิสหสัมพันธ์จะมีคุณค่าต่ำกว่าสัมประสิทธิความถดถอย [2] (β_j = Partial Regression Coefficient)

นอกจาก Simple Correlation และPartial Correlation แล้วยังมีสหสัมพันธ์อีกรูปหนึ่งคือ Multiple Correlation Coefficient ซึ่งมีประโยชน์เป็นอย่างสูงในงานวิเคราะห์ความถดถอย

---

[1] ห้ามอภิปรายในรูปของสาเหตุและผล (Causual Relationship) โดยเด็ดขาดเพราะถือว่าเป็น Nonsense Correlation ทั้งนี้มิใช่เพราะครูมีรายได้สูงขึ้นสุราตราแม่โขงจึงขายดีขึ้น และมิใช่เพราะสุราตราแม่โขงขายดีขึ้นครูจึงมีรายได้สูงขึ้น แต่เป็นเพราะมีสาเหตุอื่นที่แฝงเร้นอยู่เบื้องหลัง

[2] จากสมการ $Y = \beta_1 + \beta_2 X + u$ เราพบว่า $\hat{\beta}_2 = \sum xy / \sum x^2$ ดังนี้

$$\hat{\beta}_2 = \sum xy / \sum x^2 = \sum xy / \sum x^2 \cdot \sqrt{\sum y^2 / \sum y^2} = r s_y / s_x$$

แสดงว่า $\hat{\beta}_2$ สามารถแสดงให้เห็นถึงดีกรีความสัมพันธ์ระหว่างตัวแปรได้เช่นเดียวกับ r ขณะเดียวกับจากสมการ $E(Y) = \beta_1 + \beta_2 X$ แสดงว่า β_2 คือดัชนีที่แสดงปริมาณของอิทธิพลที่ตัวแปร X มีต่อตัวแปร Y ซึ่งย่อมแสดงว่า β_2 ซึ่งเป็น Partial Regression Coef สามารถแสดงถึงทั้งดีกรีของความสัมพันธ์และเหตุของความผันแปรในตัวแปร Y ขณะที่ r แสดงได้เพียงเฉพาะดีกรีของความสัมพันธ์เท่านั้น

ในทางทฤษฎีเรานิยาม Multiple Correlation ดังนี้[1]

$$R = \frac{\sqrt{\sigma_{11} - \sigma_{11\cdot 23\ldots k}}}{\sigma_{11}} \qquad R^2 = \frac{\sigma_{11} - \sigma_{11\cdot 23\ldots k}}{\sigma_{11}}$$

เมื่อ σ_{11} = ความผันแปรของตัวแปรสุ่ม Y โดยไม่สนใจที่จะควบคุมอิทธิพลของ $X_2, X_3, \ldots, X_k$[2] (marginal pdf ของ $Y : f_Y(y)$)

$\sigma_{11\cdot 23\ldots k}$ = ความผันแปรของตัวแปรสุ่ม Y ที่ควบคุมอิทธิพลของ $X_2, X_3, \ldots, X_k$ เอาไว้ (Conditional pdf ของ $Y : f_Y(y|x_2, x_3, \ldots, x_k)$

$\sigma_{11} - \sigma_{11\cdot 23\ldots, k}$ = ความผันแปรของตัวแปรสุ่ม Y ที่ลดลงอันสืบเนื่องมาจากการควบคุมอิทธิพลของ $X_2, X_3, \ldots, X_k$

ดังนั้น $\frac{\sigma_{11} - \sigma_{11\cdot 23 \ldots .k}}{\sigma_{11}} = R^2$ = อัตราความผันแปรของตัวแปรสุ่ม Y ที่ลดลงเนื่องจากการควบคุมอิทธิพลของตัวแปรสุ่ม $X_2, X_3, \ldots, X_k$

จะเห็นได้ว่า R^2 แสดงให้เห็นถึงปริมาณของอัตราความผันแปรของตัวแปร Y ที่ลดลงเนื่องจากการควบคุมความผันแปรในค่าของตัวแปรอิสระ ขอให้สังเกตว่า R^2 ยิ่งมีค่าสูงเพียงใดแสดงว่าปริมาณความผันแปรของ Y จะยิ่งลดลงมากเพราะสาเหตุจากการควบคุม X_j และเมื่อ $R^2 = 1$ แสดงว่า $\sigma_{11} = \sigma_{11\cdot 23\ldots k}$ หรือนัยหนึ่งการควบคุมค่าของ $X_2, X_3, \ldots, X_k$ จะมีผลให้ความผันแปรในค่าของตัวแปร Y ลดลงทั้งหมดโดยสิ้นเชิง

---

[1] ในทางปฏิบัติเรานิยามว่า R คือ Simple Correlation ระหว่าง Y และ $\hat{Y}$ เมื่อ $\hat{Y} = X\hat{\beta}$ และ Y คือเวคเตอร์ของค่าสังเกต Y

[2] ตัวอย่างเช่น Y = ผลผลิตข้าวโพด X_2 = ปริมาณน้ำฝน (ความชื้น) X_3 = อุณหภูมิ ดังนั้น σ_{11} = ความผันแปรในปริมาณผลผลิตข้าวโพดโดยมิได้สนใจว่าอุณหภูมิและความชื้นจะเปลี่ยนแปลงไปหรือไม่เพียงใด หรือนัยหนึ่ง σ_{11} คือความผันแปรในปริมาณผลผลิตข้าวโพดในทุกระดับของอุณหภูมิและความชื้น

$\sigma_{11\cdot 23}$ เมื่อ $X_2 = 10$ และ $X_3 = 30$ หมายความว่า ความผันแปรในปริมาณผลผลิตข้าวโพดที่ศึกษา ณ. ระดับความชื้น (ปริมาณน้ำฝน) 10 นิ้ว/ปี และอุณหภูมิเฉลี่ย 30 องศาเซลเซียส

$\sigma_{11} - \sigma_{11\cdot 23}$ = ความผันแปรในปริมาณผลผลิตข้าวโพดที่ลดลงภายหลังเมื่อได้ควบคุมความชื้นไว้ ณ ระดับ 10 นิ้ว/ปี และอุณหภูมิเฉลี่ย ณ ระดับ 30 องศาเซลเซียส

อย่างไรก็ตามในทางปฏิบัติเรานิยมตีความหมายของ R^2 ในรูปอีกรูปหนึ่ง ซึ่งทำความเข้าใจได้ง่ายกว่าความหมายเดิมดังนี้

R^2 หมายถึงอัตราความผันแปรของตัวแปร Y ที่ลดลงเนื่องจากอิทธิพลของตัวแปร X's ตามความหมายนี้ค่าของ R^2 จึงหมายถึงปริมาณความสามารถที่ตัวแปรอิสระสามารถอธิบายความเคลื่อนไหวของค่าตัวแปร Y นั่นเอง

4.2 เงื่อนไขและข้อจำกัดการใช้สหสัมพันธ์ในงานวิเคราะห์ความถดถอย

ในทางทฤษฎีความน่าจะเป็นเรานิยามสหสัมพันธ์เชิงเส้นดังนี้ คือ

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\{V(X)\,V(Y)\}^{1/2}} = \frac{E(X-\mu_x)(Y-\mu_y)}{\sigma_x\,\sigma_y}$$

โดยที่ $E(X-\mu_x)(Y-\mu_y) = E(XY) - \mu_x\mu_y$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dxdy - \mu_x\mu_y$$

เมื่อ $\mu_x = E(X) = \int_{-\infty}^{\infty} x\left\{\int_{-\infty}^{\infty} f(x,y)\,dy\right\}dx$

$$\sigma_x^2 = V(X) = \int_{-\infty}^{\infty}(x-\mu_x)^2\left\{\int_{-\infty}^{\infty} f(x,y)\,dy\right\}dx$$

$$\mu_y = E(Y) = \int_{-\infty}^{\infty} y\left\{\int_{-\infty}^{\infty} f(x,y)\,dx\right\}dy$$

$$\sigma_y^2 = V(Y) = \int_{-\infty}^{\infty}(y-\mu_y)^2\left\{\int_{-\infty}^{\infty} f(x,y)\,dx\right\}dy$$

เมื่อ f (x,y) คือ Joint pdf ของตัวแปรสุ่ม X และ Y

ความจริงข้อนี้ชี้ให้เห็นว่าค่าสหสัมพันธ์นั้นเป็นพารามิเตอร์ที่นิยามเฉพาะเมื่อ **X และ Y** เป็นตัวแปรสุ่ม ถ้า X และ Y มิใช่ตัวแปรสุ่มด้วยกันทั้งคู่หรือตัวแปรตัวใดตัวหนึ่งมิใช่ตัวแปรสุ่ม ค่าสหสัมพันธ์จะไม่เกิดขึ้น

พิจารณาแบบจำลอง $Y_i = \beta_1 + \sum_{j=2}^{k}\beta_j X_{ji} + u_i$; = 1, 2, ..., n จะพบว่าเราตกลงหรือถือเป็นข้อกำหนดว่า $X = (X_1, X_2, \ldots\ldots, X_k)$ คือ Fixed Matrix หรือ X_j คือ Mathemmatical Variable มิใช่ Random Variable ขณะที่ u_i เป็น Random Variable และ Y_i เป็น Random Variable เมื่อเป็นเช่นนี้ ρ_{yx_j}; j = 2, 3,...,k และ $\rho_{x_s x_j}$; s, j = 2, 3,...,k จึงปรากฏขึ้นไม่ได้ในทำนองเดียวกันค่าประมาณของ ρ คือ r_{yx_j} และ $r_{x_s x_j}$ ย่อมปรากฏขึ้นไม่ได้เช่นกัน โดยที่

$$r_{yx_j} = \Sigma (X_{ji} - \bar{X}_j)(Y_i - \bar{Y}) / \left\{ \Sigma (X_{ij} - \bar{X}_j)^2 \Sigma (Y_i - \bar{Y})^2 \right\}^{\frac{1}{2}} ; j = 2, 3, \ldots, k$$

และ $$r_{x_s x_j} = \Sigma (X_{si} - \bar{X}_s)(X_{ji} - \bar{X}_j) / \left\{ \Sigma (X_{si} - \bar{X}_s)^2 \Sigma (X_{ji} - \bar{X}_j)^2 \right\}^{\frac{1}{2}} ; s, j = 2, 3, \ldots, k$$

แต่เนื่องจากสหสัมพันธ์มีความสำคัญในงานวิเคราะห์ความถดถอยเป็นอย่างมากโดยเฉพาะอย่างยิ่งในเรื่องราวการเลือกตัวแปรอิสระสำหรับ Best Fitted Model ในกรณีของ Stepwise Regression การใช้ประโยชน์ในการประมาณค่า β การใช้ประโยชน์ในการวิเคราะห์ตาราง ANOVA การใช้ประโยชน์ในการแสดงคุณภาพของสมการถดถอย (R^2) และการใช้ประโยชน์ในการวิเคราะห์ปัญหา Autocorrelation และ Multicolinearity ความจำเป็นที่จะต้องนำสหสัมพันธ์มาใช้ในงานวิเคราะห์ความถดถอยจึงต้องมีอยู่แม้ว่า X_j จะมิใช่ตัวแปรสุ่มและมิอาจถือว่าค่าสังเกต $Z_i = (Y_i, X_{2i}, X_{3i}, \ldots, X_{ki}): i = 1, 2, \ldots, n$ เป็นค่าของตัวแปรสุ่มก็ตาม

โดยหลักเกณฑ์ทางปฏิบัติ เมื่อ $Y, X_1, X_2, \ldots, X_k$ มิใช่ตัวแปรสุ่มทั้งหมด หรือมีเฉพาะตัวแปรตัวใดตัวหนึ่งเป็นตัวแปรสุ่ม เราจะอนุโลมให้สามารถวิเคราะห์หาค่าสหสัมพันธ์ตามวิธีของ Pearson Product Moment ได้โดยจะยึดถือว่าค่า r เป็นเพียงดัชนีที่บ่งบอกความสัมพันธ์ระหว่างกลุ่มของค่าสังเกตเท่านั้น และจะไม่ยอมให้มีการอนุมานใด ๆ ในฐานะเดียวกันกับเมื่อตัวแปรเหล่านั้นเป็นตัวแปรสุ่ม[1] กล่าวคือ

1. เมื่อ X_{ji} และ X_{si} ; i = 1, 2, ... n เป็นค่าคงที่ (มิใช่กลุ่มตัวอย่างจาก pdf หนึ่งใด) หรือ X_j และ X_s เป็น Mathematical Variable กรณีนี้เราถือว่า $r_{x_j x_s}$ เป็นเพียงดัชนีวัดความสัมพันธ์ระหว่างกลุ่มของค่าสังเกตทั้งสองกลุ่มเท่านั้น และถือว่ากลุ่มของค่าสังเกตทั้งสองนั้นคือ (X_{ji}, X_{si}) ; i = 1, 2, ..., n คือกลุ่มประชากรที่มีขนาดประชากรเท่ากับ n และในกรณีนี้ $r_{x_j x_s}$ ก็คือ $\rho_{x_j x_s}$ นั่นเอง[2]

2. เมื่อ Y เป็นตัวแปรสุ่มและ X_j เป็น Mathematical Variable ให้ถือว่า X_{ji} ; i = 1, 2,...,n คือค่าของหน่วยสำรวจในกลุ่มประชากรขนาด n และถือว่า X_j มีลักษณะการแจกแจงแบบตัดตอน (Discrete) การวิเคราะห์สหสัมพันธ์ให้ใช้ Σ แทน $\int$

---

[1] Graybill, F.A. Op.Cit., p. 208

[2] Drapper, N.R. and Smith, H., Op.Cit, p.34

การวิเคราะห์ในลำดับต่อไปนี้จะนำสหสัมพันธ์มาใช้โดยยึดถือเอาความหมายที่กล่าวมานี้เป็นขอบเขตในการอภิปรายผล และให้เข้าใจไว้เสมอว่าค่าสหสัมพันธ์จะบ่งบอกเพียงดีกรีของความสัมพันธ์เท่านั้น ไม่มีการอภิปรายถึงสาเหตุและผลแต่ประการใด

4.3 ทฤษฎีเกี่ยวกับสหสัมพันธ์

ทฤษฎีที่จะกล่าวถึงในลำดับต่อไปนี้เป็นทฤษฎีที่ถือว่าตัวแปรทั้งหลายที่เกี่ยวข้องเป็นตัวแปรสุ่ม แต่เนื่องจากงานของเราถือว่า X_j เป็น Mathematical Variable ดังนั้น การนำทฤษฎีเหล่านี้ไปใช้ประโยชน์จึงเป็นเพียงการอนุโลมให้ใช้เท่านั้น การอภิปรายผลจะต้องคำนึงถึงข้อจำกัดที่กล่าวแล้วในตอน 4.2 เสมอ

ทฤษฎี 4.1 ถ้าสหสัมพันธ์ระหว่างตัวแปรสุ่ม X และ Y มีค่าเท่ากับ +1 หรือ -1 แล้วตัวแปรสุ่ม X และ Y จะมีความสัมพันธ์เชิงเส้นต่อกัน

ให้ $Z_1 = \frac{X-\mu_x}{\sigma_x}, Z_2 = \frac{Y-\mu_y}{\sigma_y}$ และสมมุติว่า $\rho_{xy} = +1$

เมื่อ $\rho_{xy} = +1$

ดังนั้น $\frac{E(X-\mu_x)(Y-\mu_y)}{\sigma_x \sigma_y} = +1$ หรือ $E\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) = E(Z_1 Z_2) = 1$

ให้ $Z = Z_1 - Z_2$

ดังนั้น $V(Z) = V(Z_1) + V(Z_2) - 2E(Z_1 Z_2) = 1 + 1 - 2 = 0$ และ $E(Z) = E(Z_1) - E(Z_2) = 0$

แสดงว่าตัวแปร Z มีการแจกแจงแบบ Point Density แสดงว่า $Z = 0$[1] หรือ $Z_1 = Z_2$ นั่นคือ

$$\frac{X-\mu_x}{\sigma_x} = \frac{Y-y}{\sigma_y}$$

$$\Rightarrow \quad Y = \left(\mu_y - \frac{\sigma_y}{\sigma_x}\mu_x\right) + \frac{\sigma_y}{\sigma_x}X$$

$$= \alpha + \beta X$$

เมื่อกำหนดให้ $\alpha = \left(\mu_y - \frac{\sigma_y}{\sigma_x}\mu_x\right)$ และ $\beta = \frac{\sigma_y}{\sigma_x}$

---

[1] ตัวแปรสุ่ม Z มีค่าความแปรปรวนเท่ากับ 0 แสดงว่าค่าของ Z ทุกค่ามิได้เบี่ยงเบนไปจากค่าเฉลี่ยเลย หรือนัยหนึ่งค่าของแปรสุ่มทุกค่าจะเท่ากับค่าเฉลี่ยในที่นี้ค่าเฉลี่ย = 0 แสดงว่า Z = 0

นั่นคือถ้า $\rho_{xy} = +1$ แล้วตัวแปรสุ่ม X และ Y จะมีความสัมพันธ์เชิงเส้นต่อกัน ในทำนองเดียวกัน เราสามารถพิสูจน์ได้ว่าถ้า $\rho_{xy} = -1$ แล้วตัวแปรสุ่ม X และ Y ก็จะมีความสัมพันธ์เชิงเส้นต่อกัน

อย่างไรก็ตามถ้า $\rho_{xy} = 0$ เราจะไม่สามารถยืนยันว่าตัวแปรสุ่ม X และ Y ไม่มีความสัมพันธ์ต่อกัน เราเพียงยืนยันไว้แต่เพียงว่าตัวแปรสุ่มทั้งสองไม่มีความสัมพันธ์เชิงเส้นต่อกันเท่านั้น เพราะอาจมีความสัมพันธ์ในรูปที่มิใช่เชิงเส้นได้ นอกจากนี้ $\rho_{xy} = 0$ มิได้ยืนยันว่าตัวแปรสุ่ม X และ Y เป็นอิสระต่อกันเว้นเฉพาะเมื่อ f (x, y) มีการแจกแจงปกติเท่านั้น ตัวอย่างเช่นเรามีค่าคู่ลำดับ(X,Y) อยู่ 4 คู่ แต่ละจุดให้ความน่าจะเป็นเท่ากับ $\frac{1}{4}$ ดังนี้คือ (1, 1), (2, 4), (-1, 1), (-2, 4) ซึ่งจะสังเกตได้ว่า $Y = X^2$ และพบว่า

$$E(X) = 1\left(\tfrac{1}{4}\right) + 2\left(\tfrac{1}{4}\right) + (-1)\left(\tfrac{1}{4}\right) + (-2)\left(\tfrac{1}{4}\right) = 0$$

$$E(X^2) = 1\left(\tfrac{1}{4}\right) + 4\left(\tfrac{1}{4}\right) + 1\left(\tfrac{1}{4}\right) + 4\left(\tfrac{1}{4}\right) = \frac{10}{4}$$

$$E(Y) = 1\left(\tfrac{1}{4}\right) + 4\left(\tfrac{1}{4}\right) + 1\left(\tfrac{1}{4}\right) + 4\left(\tfrac{1}{4}\right) = \frac{10}{4}$$

$$E(Y^2) = 1\left(\tfrac{1}{4}\right) + 16\left(\tfrac{1}{4}\right) + 1\left(\tfrac{1}{4}\right) + 16\left(\tfrac{1}{4}\right) = \frac{34}{4}$$

$$E(XY) = 1\left(\tfrac{1}{4}\right) + 8\left(\tfrac{1}{4}\right) + (-1)\left(\tfrac{1}{4}\right) + (-8)\,\tfrac{1}{4} = 0$$

ดังนั้น $$\rho_{xy} = \frac{E(XY) - \mu_x \mu_y}{\sigma_x \sigma_y} = 0$$

โดยที่ $\sigma_x^2 = E(X^2) - (E(X))^2$, $\sigma_y^2 = E(Y^2) - (E(Y))^2$

จะเห็นว่า $\rho_{xy} = 0$ แต่ตัวแปรสุ่ม X และ Y มีความสัมพันธ์กันในรูป $Y = X^2$ ซึ่งยืนยันให้เห็นว่าเมื่อ $\rho_{xy} = 0$ นั้นเราจะด่วนสรุปว่าตัวแปรสุ่มไม่มีความสัมพันธ์ต่อกันมิได้ และในขณะเดียวกันเราก็ไม่อาจยืนยันว่าตัวแปรสุ่ม X และ Y เป็นอิสระต่อกันได้เช่นกัน

ทฤษฎี 4.2 เมื่อตัวแปรสุ่ม X และ Y มี Joint pdf เป็น Bivariate Normal Distribution และค่าประมาณของ ρ คือ

$$r_{xy} = \hat{\rho} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \Big/ \sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2}$$

Sampling Distribution ของ $\hat{\rho}$ จะปรากฏดังนี้

$$f(\hat{\rho},\ P) = \frac{(n-2)(1-\rho^2)^{\frac{1}{2}(n-1)}}{\pi} (1-\hat{\rho}^2)^{\frac{1}{2}(n-4)} \int_0^\infty \frac{1}{(\text{Cosh } w - \hat{\rho}\rho)^{n-1}}\, dw\,; \quad -1 \leq \rho \leq 1$$

หรือเสนอได้อีกรูปหนึ่งดังนี้[1]

$$f(\hat{\rho},\rho)=\frac{(1-\rho^2)^{\frac{1}{2}(n-1)}}{\sqrt{\pi}\,\Gamma(\frac{n-1}{2})\,\Gamma(\frac{n-2}{2})}(1-\hat{\rho}^2)^{\frac{1}{2}(n-4)}\sum_{k=0}^{\infty}\frac{(2\rho\hat{\rho})^k}{k!}\Gamma^2\frac{(n-1+k)}{2},\ -1<\hat{\rho}<1$$

และเมื่อ $n\to\infty$ จะพบว่า $\hat{\rho}\sim N[\rho,\frac{(1-\rho^2)^2}{2}]$ แต่ถ้า $n<50$ การประมาณด้วยการแจกแจงปกติจะมีข้อผิดพลาด

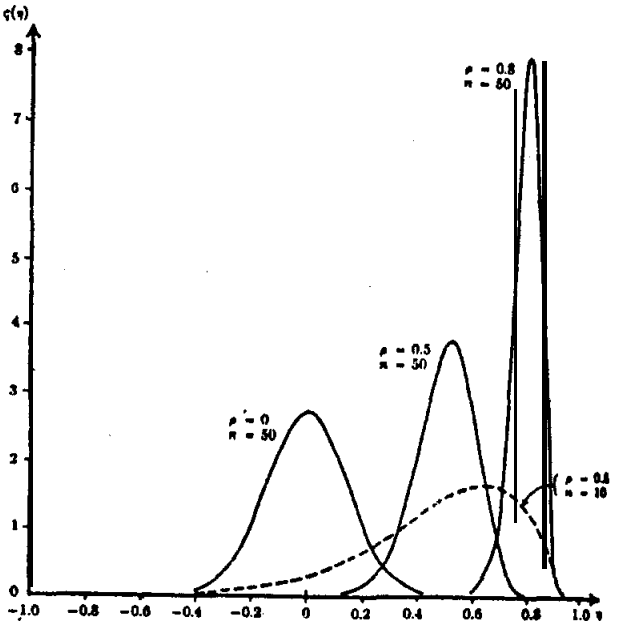
สำหรับกรณีเฉพาะเมื่อ $\rho=0$ จะปรากฏลักษณะที่น่าสนใจดังนี้

1. $f(\hat{\rho},0)=\frac{\Gamma(\frac{n-1}{2})(1-\rho^2)^{\frac{1}{2}(n-4)}}{\sqrt{\pi}\,\Gamma(\frac{n-2}{2})}$; $-1\leq\hat{\rho}\leq 1$ โดยที่ $E(\hat{\rho})=0$ และ $V(\hat{\rho})=\frac{1}{n}$

2. $f(\hat{\rho}^2,0)=\frac{\Gamma(\frac{n-1}{2})(1-\hat{\rho}^2)^{\frac{1}{2}(n-4)}}{\sqrt{\pi}\,\Gamma(\frac{n-2}{2})(\hat{\rho}^2)^{\frac{1}{2}}}$; $0<\hat{\rho}^2<1$ โดยที่ $E(\hat{\rho}^2)=\frac{1}{n-1}$

3. สำหรับการทดสอบสมมุติฐาน $H_0:\rho=0$ VS $H_1:\rho\neq 0$ จะพบว่า $t_c=\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$ มีการแจกแจงแบบ t ที่มี df $=n-2$

การแจกแจงของ $\hat{\rho}$ ปรากฏดังภาพ ขอให้สังเกตว่าโค้งจะเบี่ยงเบนลักษณะของการแจกแจงปกติถ้า n มีค่าต่ำกว่า 50



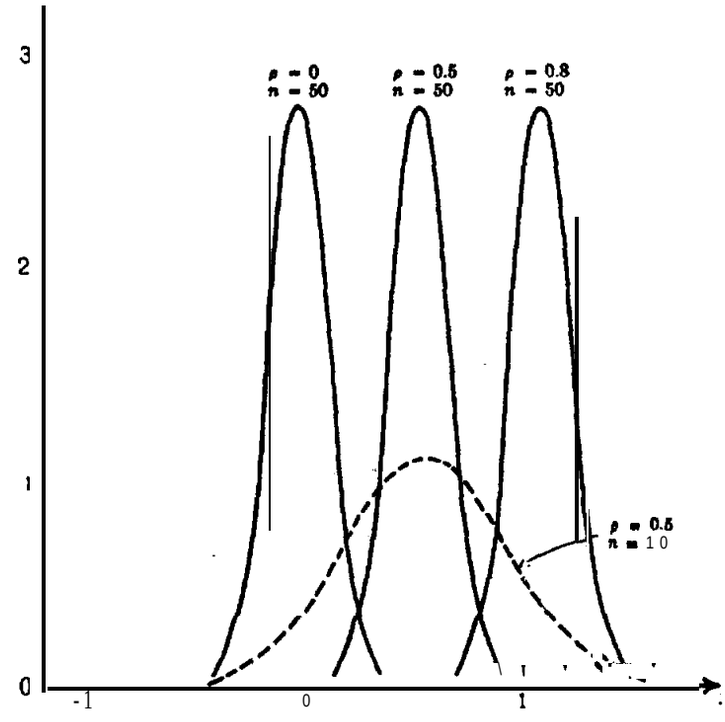
[1] $\frac{1}{\sqrt{\pi}\,\Gamma(\frac{n-1}{2})\,\Gamma(\frac{n-2}{2})}=\frac{2^{n-3}}{\pi\,\Gamma(n-2)}$, ถ้า n เป็นปริมาณติดลบ $\Gamma(n)=\frac{\Gamma(n+1)}{n}$

สำหรับการทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ ρ ในลักษณะอื่นที่ต่างไปจากลักษณะที่กล่าวมาแล้ว เช่น $H_o : \rho = \rho^*$ Vs $H_1 : \rho \neq \rho^*$ หรือ $\rho > \rho^*$ หรือ $\rho < \rho^*$ เมื่อ ρ^* คือค่าคงที่ที่กำหนดขึ้นด้วยเหตุผลใดเหตุผลหนึ่ง หรือจากประสบการณ์ในอดีต *(Priori Information)* ให้อาศัย Fisher's Transformation ซึ่งสามารถใช้สำหรับทดสอบสมมุติฐานได้ทั้งกรณี Single Parameter และ Several Parameter ดังนี้

$$\text{ให้ } Z = \frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}} \text{ และให้ } \delta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \text{ จะพบว่า}$$

ตัวแปรสุ่ม Z-δ มีการแจกแจงแบบปกติทั้งนี้เป็นจริงทั้งในกรณี n มีขนาดเล็กและขนาดใหญ่ และพบว่า $Z \sim N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)}, \frac{1}{n-3}\right)$ หรือเมื่อยอมให้มีความผิดพลาดในการประมาณค่าได้บ้างโดยไม่มีผลเสียหายมากนักจะถือว่า $Z \sim N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$ และถ้า Z_1 และ Z_2 คือค่าของ Z ที่คำนวณจากกลุ่มตัวอย่างขนาด n_1 และ n_2 ที่สุ่มมาจากกลุ่มประชากรเดียวกันแล้วเราจะพบว่า $Z_1 - Z_2 \sim N\left(0, \frac{1}{n_1-3}+\frac{1}{n_2-3}\right)$

การแจกแจงของตัวแปรสุ่ม Z ปรากฏดังภาพ ขอให้สังเกตว่าโค้งของ Z จะมีลักษณะของโค้งปกติเสมอแม้ว่า n จะมีขนาดเล็ก



ทฤษฎี 4.3 ในการทดสอบสมมุติฐานเกี่ยวกับ ρ เมื่อระบุค่า ρ ในสมมุติฐานหลักเป็นค่าคงที่ใด ๆ ที่มิใช่ 0 ให้ปฏิบัติดังนี้

1. $H_0 : \rho = \rho^*$ ถ้า $n \leq 25$ ให้ทดสอบโดยใช้ David's Table ถ้า $n \geq 25$ ให้ตัดสินใจโดยถือว่า

$$u = \sqrt{n-3}\left\{\left(\frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) - \left(\frac{1}{2}\ln\frac{1+\rho^*}{1-\rho^*}\right)\right\} \sim N(0,1)$$

2. สำหรับ $H_0 : \rho_1 = \rho_2 = \ldots \rho_k = \rho^*$ ให้ตัดสินใจโดยถือว่า

$$U = \sum_i^k (n_i-3)\left\{\left(\frac{1}{2}\ln\frac{1+\hat{\rho}_i}{1-\hat{\rho}_i}\right) - \left(\frac{1}{2}\ln\frac{1+\rho^*}{1-\rho^*}\right)\right\}^2 \sim \chi^2(k)$$

เมื่อ $\hat{\rho}_i$ คือค่าประมาณของ ρ_i ที่คำนวณโดยอาศัยข้อมูลจากกลุ่มตัวอย่างขนาด n_i ที่สุ่มจากกลุ่มประชากร Bivariate Normal กลุ่มที่ i ; $i = 1, 2, \ldots, k$

3. สำหรับ $H_0 : \rho_1 = \rho_2 = \ldots = \rho_k$ โดยมิได้ระบุค่าที่แน่ชัดของค่าสหสัมพันธ์เหล่านี้ว่าเท่ากับเท่าไรให้ตัดสินใจโดยถือว่า

$$W = \sum_i^k (n_i-3)\left\{\left(\frac{1}{2}\ln\frac{1+\hat{\rho}_i}{1-\hat{\rho}_i}\right) - \bar{Z}\right\}^2 \sim \chi^2_{(k-1)}$$

$$\text{โดยที่ } \bar{Z} = \frac{\sum_i^k (n_i-3)\left(\frac{1}{2}\ln\frac{1+\hat{\rho}_i}{1-\hat{\rho}_i}\right)}{\sum(n_i-3)}$$

สำหรับกรณีของ Partial Correlation และ Multiple Correlation นั้นเราจำเป็นต้องคำนวณโดยอาศัย Conditional Distribution เพราะ Partial Correlation คือสหสัมพันธ์ระหว่างตัวแปรสุ่มคู่ใด ๆ เมื่อกำหนดว่าตัวแปรสุ่มอื่น ๆ ที่เหลือมีอิทธิพลคงที่[1] ส่วน Multiple Correlation คือ Simple Correlation ระหว่างตัวแปรอิสระตัวใดตัวหนึ่งกับตัวแปรอื่น ๆ ที่เหลือทั้งหมด

ในที่นี้จำเป็นต้องเสนอหลักเกณฑ์ทางทฤษฎีสำหรับเรื่องนี้ไว้พอเป็นแนวทางส่วนวิธีการที่ใช้จริงในทางปฏิบัติจะได้กล่าวถึงต่อไป

ให้เวคเตอร์ Y ขนาด $p \times 1$ มีการแจกแจงแบบ $N(\mu, V)$ และเรา Partition เวคเตอร์ Y ออกเป็น 2 ส่วนคือ $Y_{(1)}$ ขนาด $q \times 1$ และ $Y_{(2)}$ ขนาด $(p-q) \times 1$ นั่นคือ $Y = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \end{bmatrix}$

---

[1] เช่น $\rho_{13.567}$ ก็คือสหสัมพันธ์ระหว่าง Y_1 และ Y_3 ใน $f(Y_1, Y_2, Y_3, Y_4 | Y_5, Y_6, Y_7)$

และถ้าหากว่า $Y^* = \begin{bmatrix} Y_{(1)} \\ Y^*_{(2)} \end{bmatrix}$, $\mu = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$, $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$ เป็นเวคเตอร์และเมตริกซ์ที่ถูกแบ่งออกเป็นส่วน ๆ ให้สอดคล้องกันในลักษณะที่การคูณและการบวกแมตริกซ์และเวคเตอร์เป็นไปได้ตามนิยาม

ดังนั้น Conditional Distribution ของ $Y_{(1)}$ เมื่อควบคุมให้ $Y_{(2)} = Y^*_{(2)}$ จะมีการแจกแจง แบบ Multivariate Normal โดยที่

$$E(Y_{(1)} \mid Y^*_{(2)}) = U_1 + V_{12} V_{22}^{-1} (Y^*_{(2)} - U_2)$$

และ $$V(Y_{(1)} \mid Y^*_{(2)}) = V_{11} - V_{12} V_{22}^{-1} V_{21}$$

$$f(Y_{(1)} \mid Y_{(2)}^*) =$$

$$\frac{1}{K^*} \exp[-\tfrac{1}{2}\{Y_{(1)} - U_1 - V_{12} V_{22}^{-1}(Y^*_{(2)} - U_2)\}' \{ V_{11} - V_{12} V_{22}^{-1} V_{21}\}^{-1}\{Y_{(1)} - U_1 - V_{12} V_{22}^{-1} (Y^*_{(2)} - U_2)\}]$$

โดยที่ $K^* = (2\pi)^{q/2} |V_{11} - V_{12} V_{22}^{-1} V_{21}|^{\frac{1}{2}}$

เมื่อเราให้ชื่อของ $V(Y_{(1)} \mid Y^*_{(2)})$ เสียใหม่เป็น $V_{11\cdot 2}$ คือ $V_{11\cdot 2} = V_{11} - V_{12} V_{22}^{-1} V_{21}$ เราจะสามารถหา Partial Correlation ระหว่าง Y_i และ Y_j (Y_i และ Y_j เป็นเวคเตอร์ ใน $Y_{(1)}$ เมื่อถือว่าเวคเตอร์ใน $Y_{(2)}$ มีอิทธิพลคงที่เท่ากับ $Y^*_{(2)}$ ได้ดังนี้

$$\rho_{ij \cdot q+1,\, q+2, \ldots, p} = \frac{\sigma_{ij \cdot q+1,\, q+2, \ldots, p}}{\sqrt{\sigma_{ii \cdot q+1, \ldots, p}\, \sigma_{jj \cdot q+1, \ldots, p}}}$$

โดยที่ $\sigma_{ij \cdot q+1, \ldots, p}$ คือสมาชิกลำดับที่ (i, j) ใด ๆ ในเมตริกซ์ $V_{11\cdot 2}$

ในขณะเดียวกัน ถ้าเรา Partition เวคเตอร์ Y เป็น $Y = (Y_{(1)}, Y_{(2)})'$ โดยที่ $Y_{(1)}$ คือเวคเตอร์ขนาด 1×1 หรือ $Y = (Y_1 \mid Y_2, Y_3, \ldots, Y_p)' = (Y_{(1)} \mid Y_{(2)})'$ และPartial เวคเตอร์ μ และเมตริกซ์ V ดังนี้คือ

$$\mu = (\mu_1 \mid \mu_2, \mu_3, \ldots, \mu_p)' = (U_1 \mid U_2)' \quad , V = \left[\begin{array}{c|ccc} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \hline \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \cdots & \sigma_{3p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{array}\right] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

ให้ $Z = V_{12} V_{22}^{-1}(Y_{(2)}-U_2)$

ดังนั้น Multiple Correlation คือสหสัมพันธ์ระหว่างตัวแปรสุ่ม Y_1 และ Z คือ

$$P_{\cdot} = \frac{E\{(Y_1-\mu_1)(Z-E(Z))\}}{\sqrt{E(Y_1-\mu_1)^2 E(Z-E(Z))^2}}$$

$$= \sqrt{\frac{V_{12} V_{22}^{-1} V_{21}}{\sigma_{11}}} = \sqrt{\frac{\sigma_{11}-\sigma_{11\cdot 23\cdots p}}{\sigma_{11}}}$$

ตัวอย่าง 4.1 เวคเตอร์ Y ขนาด 4×1มีการแจกแจงปกติ $N(\mu, V)$ โดยที่

$$\mu = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -1 \end{bmatrix} \quad \text{และ} \quad V = \begin{bmatrix} 2/5 & -1/5 & 0 & 0 \\ -1/5 & 3/5 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

จงคำนวณ $f(Y_1 | Y_2, Y_3, Y_4)$, $f(Y_1)$, ρ_{12}, $\rho_{12\cdot 4}$ และ ρ_1

วิธีทำ

ก $f(Y_1 | Y_2, Y_3, Y_4) = ?$

เราทราบว่า $f(Y_{(1)} | Y_{(2)})$ มีการแจกแจงแบบปกติโดยที่

$$E(Y_{(1)} | Y_{(2)}) = U_1 + V_{12} V_{22}^{-1} (Y_{(2)} - U_2) \text{ และ } V(Y_{(1)} | Y_{(2)}) = V_{11} - V_{12} V_{22}^{-1} V_{21} = \sigma_{11\cdot 234}$$

จากเวคเตอร์ μ และเมตริกซ์ V เราสามารถแยกเป็นแมตริกซ์ย่อยที่สอดคล้องกับนิยามการบวกและการคูณได้ดังนี้

$$\mu = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad V = \left[\begin{array}{c|cc} 2/5 & -1/5 & 0 \\ \hline -1/5 & 3/5 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{array}\right] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

$$\text{และพบว่า } V_{22}^{-1} = \begin{bmatrix} 3/5 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 5/3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

ดังนั้น $E(Y_{(1)} \mid Y_{(2)}) = U_1 + V_{12} V_{22}^{-1} (Y_{(2)} - U_2) = 1 + (-1/5, 0, 0) \begin{bmatrix} 5/3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} Y_2 - 0 \\ Y_3 - 2 \\ Y_4 - (-1) \end{bmatrix}$

$= 1 - \frac{1}{3} Y_2$

$V(Y_{(1)} \mid Y_{(2)}) = V_{11} - V_{12} V_{22}^{-1} V_{21} = \sigma_{11 \cdot 234}$ หรือ $V_{11 \cdot 2} = \frac{2}{5} - (-\frac{1}{5}, 0, 0) \begin{bmatrix} 5/3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{5} \\ 0 \\ 0 \end{bmatrix}$

$= \frac{2}{5} - \frac{1}{15} = \frac{1}{3}$

นั่นคือ $f(Y_1 \mid Y_2, Y_3, Y_4) = N(1 - \frac{1}{3} Y_2, \frac{1}{3})$

ข. $f(Y_1) = ?$

จากเมตริกซ์ μ และ V ที่แบ่งเป็นเมตริกซ์ย่อยแล้วในข้อ ก พบว่า $Y_1 \sim N(1, 2/5)$

ค. สหสัมพันธ์ระหว่าง Y_1 และ Y_2 คือ ρ_{12} มีค่าเท่ากับเท่าไร

จากนิยาม $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ โดยพิจารณา $f(Y_i, Y_j)$

ดังนั้น $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} = \frac{-1/5}{\sqrt{(2/5)(3/5)}} = \frac{1}{\sqrt{6}}$

ง. $\rho_{12 \cdot 4} = ?$

$\rho_{12 \cdot 4}$ คือสหสัมพันธ์ระหว่าง Y_1 และ Y_2 เมื่อถือว่า Y_4 มีอิทธิพลคงที่จะเห็นว่า ค่า $\rho_{12 \cdot 4}$ สามารถคำนวณได้จาก $f(Y_1, Y_2, Y_3 \mid Y_4)$

เราพบว่า $E(Y_{(1)} \mid Y_{(2)}) = U_1 + V_{12} V_{22}^{-1} (Y_{(2)} - U_2)$ และ $V(Y_{(1)} \mid Y_{(2)}) = V_{11} - V_{12} V_{22}^{-1} V_{21}$

โดยที่แมตริกซ์ V ได้รับการแบ่งเป็นแมตริกซ์ย่อยดังนี้

$$V = \left[\begin{array}{ccc:c} 2/5 & -1/5 & 0 & 0 \\ -1/5 & 3/5 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ \hdashline 0 & 0 & -1 & 2 \end{array}\right] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

ดังนั้น $V_{11} - V_{12}V_{22}^{-1}V_{21} = \begin{bmatrix} 2/5 & -1/5 & 0 \\ -1/5 & 3/5 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} (1/2)(0, 0, -1) = \begin{bmatrix} 2/5 & -1/5 & 0 \\ -1/5 & 3/5 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$

ดังนั้น $\rho_{12\cdot4} = \dfrac{\sigma_{12\cdot4}}{\sqrt{\sigma_{11\cdot4}\sigma_{22\cdot4}}} = \dfrac{-1/5}{\sqrt{(2/5)(3/5)}} = \dfrac{1}{\sqrt{6}}$

จ. $\rho_1 = ?$

จาก $f(Y_1|Y_2, Y_3, Y_4)$ ในข้อ ก. ดังนั้น

$$\rho_1 = \sqrt{\frac{V_{12}V_{22}^{-1}V_{21}}{V_{11}}} = \sqrt{\frac{1/15}{2/5}} = \frac{1}{\sqrt{6}}$$

ดังนั้น $\rho_1^2 = 1/6$

หรือ $\because \sigma_{11} = V_{11} = 2/5$ และจากข้อ ก. $\sigma_{11\cdot234} = 1/3$ ดังนั้น

$$\rho_1^2 = \frac{1-\sigma_{11\cdot234}}{\sigma_{11}} = \frac{1-1/3}{2/5} = 1-\frac{5}{6} = 1/6$$

สำหรับในทางปฏิบัตินั้นเราไม่อาจทราบค่าจริงของ V ได้ดังตัวอย่างข้างต้น ดังนั้นเราจำเป็นต้องอาศัยกระบวนการสุ่มตัวอย่างเวคเตอร์ของค่าสังเกต $Z_i = (Y_{1i}, Y_{2i}, \ldots\ldots, Y_{pi})$; $i = 1, 2, \ldots, n$ มา n ชุดแล้วคำนวณหาเมตริกซ์ B โดยที่ $B = \sum_{i}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})' = (n-1)S$ เมื่อ S คือตัวประมาณค่าของ V

จากนั้นให้แยกแมตริกซ์ B เป็นเมตริกซ์ย่อยทำนองเดียวกันกับการแบ่งแมตริกซ์ V และคำนวณหาค่า Partial Correlation โดยอาศัยหลักเกณฑ์ทำนองเดียวกันกับเมื่อกระทำกับเมตริกซ์ V กล่าวคือ

จำแนกเมตริกซ์ B ทำนองเดียวกับ V ดังนี้คือ

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

และให้ชื่อเมตริกซ์ $B_{11} - B_{12}B_{22}^{-1}B_{21}$ เป็น $B_{11\cdot2}$ และให้ $b_{ij\cdot2}$ หมายถึงสมาชิกที่ (ij) ของ $B_{11\cdot2}$

ดังนั้น $\hat{\rho}_{ij\cdot q+1, q+2, \ldots, p} = \dfrac{b_{ij\cdot2}}{\sqrt{b_{ii\cdot2}b_{jj\cdot2}}}$

ทั้งนี้เราสามารถพิสูจน์ได้ว่า $b_{ij\cdot 2} \sim$ Wishart $[p, (n-1) - (p-q), V_{11\cdot 2}]$[1] ส่วนงานทดสอบสมมุติฐานให้อาศัยทฤษฎี 3.2 และ 3.3 โดยให้ใช้ $(n-p+q)$ แทนที่ของ n และใช้ $\rho_{ij\cdot q+1,\, q+2,\ldots,\, p}$ ในที่ของ ρ

ตัวอย่าง 4.2 สุ่มตัวอย่าง Random Vector Y_i ขนาด 4×1 มา 24 ชุด ปรากฏเมตริกซ์ B ดังนี้

$$B = \sum_{i}^{24} \begin{bmatrix} Y_{i1} - \bar{Y}_i \\ Y_{i2} - \bar{Y}_i \\ Y_{i3} - \bar{Y}_i \\ Y_{i4} - \bar{Y}_j \end{bmatrix} (Y_{i1}-\bar{Y}_i,\ Y_{i2}-\bar{Y}_i,\ Y_{i3}-\bar{Y}_i,\ Y_{i4}-\bar{Y}_i)$$

$$= \begin{bmatrix} 10 & 9 & -1 & -16 \\ 9 & 20 & -3 & -16 \\ -1 & -3 & 5 & 3 \\ -16 & -16 & 3 & 27 \end{bmatrix}$$

จงหาค่า $\hat{\rho}_{12\cdot 34}$

---

[1] ทฤษฎีเกี่ยวกับ Wishart Distribution ปรากฏดังนี้

ถ้า $Y \sim N[0, V]$ โดยที่ Y คือเวคเตอร์ขนาด $p\times 1$ และให้ Y_j, $j = 1, 2, \ldots, n$ $(n \geq p)$ เป็น Random Vector ที่สุ่มมาจากประชากรดังกล่าว แล้วสมาชิกใด ๆ ของเมตริกซ์ $B = \sum_{j}^{n} Y_j Y_j'$ จะมี Joint pdf ดังนี้

$$f(b_{11}, b_{12}, \ldots, b_{pp}) = \begin{cases} \dfrac{\exp\{-\frac{1}{2}\mathrm{tr}(BV^{-1})\}\ |B|^{\frac{1}{2}(n-p-1)}}{\pi^{\frac{1}{4}p(p-1)}\, 2^{np/2} \prod_{i}^{p} \Gamma\left(\frac{n+1-i}{2}\right)} & \text{ถ้า B เป็น positive definite} \\ 0 & \text{ถ้า B ไม่เป็น positive definite} \end{cases}$$

ดังนั้นเมตริกซ์ B จะมีการแจกแจงแบบ Wishart คือ $W(p, n, V)$ และถ้า Y_j เป็น Random Vector ที่สุ่มจากกลุ่มประชากร $N(\mu, V)$ แล้ว $B = \sum^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'$ จะมีการแจกแจง $W(p, n-1, V)$

วิธีทำ แบ่งแมตริกซ์ B เป็นเมตริกซ์ย่อยดังนี้

$$B = \left[\begin{array}{cc:cc} 10 & 9 & -1 & -16 \\ 9 & 20 & -3 & -16 \\ \hdashline -1 & -3 & 5 & 3 \\ -16 & -16 & 3 & 27 \end{array}\right] = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$\text{ดังนั้น } B_{11\cdot 2} = B_{11} - B_{12}B_{22}^{-1}B_{21} = \begin{bmatrix} 10 & 9 \\ 9 & 20 \end{bmatrix} - \begin{bmatrix} -1 & -16 \\ -3 & -16 \end{bmatrix}\begin{bmatrix} 5 & 3 \\ 3 & 27 \end{bmatrix}^{-1}\begin{bmatrix} -1 & -3 \\ -16 & -16 \end{bmatrix}$$

$$= \frac{1}{126}\begin{bmatrix} 49 & -35 \\ -35 & 1{,}285 \end{bmatrix} = \begin{bmatrix} .389 & -.278 \\ -.278 & 10.198 \end{bmatrix}$$

$$\text{ดังนั้น } \quad \hat{\rho}_{12\cdot 34} = \frac{b_{12\cdot 34}}{\sqrt{b_{11\cdot 34}\, b_{22\cdot 34}}} = \frac{-0.278}{\sqrt{(0.389)(10.198)}} = -0.1395$$

4.4 การวิเคราะห์ความถดถอยโดยอาศัยสหสัมพันธ์

4.4.1 วิธีการและการพัฒนา

การหาแมตริกซ์ในตอน 4.3 มีลักษณะที่เชื่อมโยงนำไปสู่การคำนวณหา Correlation Matrix ซึ่งสามารถใช้ในการคำนวณหา Partial Correlation, Multiple Correlation และการนำไปใช้ประโยชน์ในงานทดสอบสมมุติฐานในรูป ANOVA

ก่อนอื่นขอให้นักศึกษาพิจารณาสมการถดถอยต่อไปนี้

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \; ; i = 1, 2, \ldots, n \qquad \ldots\ldots\ldots (1)$$

จาก (1) ถ้าเรารวมตลอดในทุกค่าของ i และหารตลอดด้วย n จะพบว่า

$$\bar{Y} = \beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \ldots + \beta_k \bar{X}_k + \bar{u} \qquad \ldots\ldots\ldots (2)$$

$$(1) - (2) \quad Y_i - \bar{Y} = \beta_2 (X_{2i} - \bar{X}_2) + \beta_3 (X_{3i} - \bar{X}_3) + \ldots + \beta_k (X_{ki} - \bar{X}_k) + (u_i - \bar{u})$$

$$; i = 1, 2, \ldots, n \qquad \ldots\ldots\ldots (3)$$

ถ้าให้ $Y_i - \bar{Y} = y_i$ และ $X_{ji} - \bar{X}_j = x_{ji}$ สมการถดถอยจะเปลี่ยนจากรูปที่ต้องวิเคราะห์โดยอาศัยข้อมูลเดิม (Initial data) เป็นรูปที่วิเคราะห์โดยอาศัยข้อมูลที่แปลงรูปแล้ว (Deviated Form) สมการถดถอย (1) จึงเปลี่ยนรูปไปสู่สมการ (4) ดังนี้

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + (u_i - \bar{u}) \; ; \; i = 1,2,\ldots,n \qquad \ldots\ldots\ldots(4)$$

ขอให้สังเกตสมการ (4) จะปรากฏเฉพาะสัมประสิทธิความถดถอย $\beta_2, \beta_3, \ldots, \beta_k$ เท่านั้น ส่วน β_1 ซึ่งมิใช่สัมประสิทธิความถดถอยจะไม่ปรากฏอยู่ นอกจากนี้ข้อมูลที่เสนอในรูป Deviated Form ก็คือข้อมูลที่เกิดจากการย้ายแกนของตัวแปรไปยังแกนใหม่ซึ่งมีจุดกำเนิด ณ ค่าเฉลี่ยซึ่งข้อมูลลักษณะนี้จะอำนวยประโยชน์ในการวิเคราะห์สหสัมพันธ์ได้เป็นอย่างดี

พิจารณาสมการ $y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + (u_i - \bar{u}); \; i = 1, 2, \ldots, n$

จัดเป็นรูปเมตริกซ์

$$\Rightarrow Y = X\beta + U - \bar{U}$$

ให้ $\hat{\beta}$ คือ OLS - Estimator ของ β

$$\Rightarrow Y = X\hat{\beta} + e$$

โดยที่ $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} x_{21} & x_{31} & \cdots & x_{k1} \\ x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{2n} & x_{3n} & \cdots & x_{kn} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix}$, $U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$, $\bar{U} = \begin{bmatrix} \bar{u} \\ \bar{u} \\ \vdots \\ \bar{u} \end{bmatrix}$

โดยอาศัยหลักเกณฑ์ของ OLS เราสามารถพิสูจน์ได้ว่า $\hat{\beta} = (X'X)^{-1}X'Y$ และ $\hat{V}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$ โดยที่ $\hat{\sigma}^2 = \frac{1}{n-k}(Y'Y - \hat{\beta}'X'Y)$ เมื่อ k = จำนวนพารามิเตอร์ β และพบว่า $E(\hat{\beta}) = \beta$ ซึ่งสามารถแสดงให้เห็นได้ดังนี้

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

แต่ $X'\bar{U} = \begin{bmatrix} X_{21} & X_{22} & \dots & X_{2n} \\ X_{31} & X_{32} & \dots & X_{3n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} \bar{u} \\ \bar{u} \\ \bar{u} \\ \vdots \\ \bar{u} \end{bmatrix} = \begin{bmatrix} \bar{u}\Sigma x_{2i} \\ \bar{u}\Sigma x_{3i} \\ \vdots \\ \bar{u}\Sigma x_{ki} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

ดังนั้น $\hat{\beta} = \beta + (X'X)^{-1}X'U$ และ $E(\hat{\beta}) = \beta$ (เพราะ $E(U) = 0$) และพบว่าสมการถดถอย คือ

$$E(y) = \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k$$

หรือ $y = \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k$

สำหรับการคำนวณหา Correlation Table ให้กระทำดังนี้

จากสมการ Normal คือ $(X'X)\hat{\beta} = X'Y$ จะพบว่า

$$\begin{bmatrix} \sum_i^n x_{2i}^2 & \sum_i^n x_{2i}x_{3i} & \dots & \sum_i^n x_{2i}x_{ki} \\ \sum_i^n x_{2i}x_{3i} & \sum_i^n x_{3i}^2 & \dots & \sum_i^n x_{3i}x_{ki} \\ \vdots & & & \\ \sum_i^n x_{2i}x_{ki} & \sum_i^n x_{3i}x_{ki} & \dots & \sum_i^n x_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \\ \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \Sigma x_{2i}y_i \\ \Sigma x_{3i}y_i \\ \\ \Sigma x_{ki}y_i \end{bmatrix}$$

หรือ

$$\begin{aligned} \hat{\beta}_2 \Sigma x_2^2 + \hat{\beta}_3 \Sigma x_2x_3 + \dots + \hat{\beta}_k \Sigma x_2x_k &= \Sigma x_2y \\ \hat{\beta}_2 \Sigma x_2x_3 + \hat{\beta}_3 \Sigma x_3^2 + \dots + \hat{\beta}_k \Sigma x_3x_k &= \Sigma x_3y \qquad \dots\dots (1) \\ &\vdots \\ \hat{\beta}_2 \Sigma x_2x_k + \hat{\beta}_3 \Sigma x_3x_k + \dots + \hat{\beta}_k \Sigma x_k^2 &= \Sigma x_ky \end{aligned}$$

แต่ $r_{ij} = \dfrac{\Sigma x_i x_j}{\sqrt{\Sigma x_i^2 \, \Sigma x_j^2}}$; i, j = 2, 3, ... k โดยที่ $r_{ij} = 1$ ถ้า i = j

ดังนั้น $\Sigma x_i x_j = r_{ij} \sqrt{\Sigma x_i^2} \sqrt{\Sigma x_j^2} = (n-1) r_{ij} s_i s_j$ เมื่อ $s_i^2 = \dfrac{1}{n-1} \Sigma x_{it}^2$

ดังนั้นระบบสมการ (1) จึงเปลี่ยนรูปเป็น

$$\hat{\beta}_2 r_{22} s_2 + \hat{\beta}_3 r_{23} s_3 + \ldots + \hat{\beta}_k r_{2k} s_k = r_{12} s_1$$
$$\hat{\beta}_2 r_{23} s_2 + \hat{\beta}_3 r_{33} s_3 + \ldots + \hat{\beta}_k r_{3k} s_k = r_{13} s_1 \qquad \ldots(2)$$

$$\hat{\beta}_2 r_{2k} s_2 + \hat{\beta}_3 r_{3k} s_3 + \ldots + \hat{\beta}_k r_{kk} s_k = r_{1k} s_1$$

ดังนั้น
$$\begin{bmatrix} r_{22} & r_{23} & \cdots & r_{2k} \\ r_{23} & r_{33} & \cdots & r_{3k} \\ \vdots & \vdots & & \vdots \\ r_{2k} & r_{3k} & \cdots & r_{kk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_2 s_2 \\ \hat{\beta}_3 s_3 \\ \vdots \\ \hat{\beta}_k s_k \end{bmatrix} = \begin{bmatrix} r_{12} s_1 \\ r_{13} s_1 \\ \vdots \\ r_{1k} s_1 \end{bmatrix} \text{ หรือ } R_1 B^* = G \qquad \ldots(3)$$

ซึ่งจากระบบสมการ $R_1 B^* = G$ เราจะพบว่า R_1 คือ Correlation Matrix ซึ่งการคำนวณหาค่า $\hat{\beta}_i$; $i = 2, 3, \ldots, k$ สามารถกระทำได้หลายวิธีเช่นคำนวณหา $\hat{\beta}_i$ จากเวคเตอร์ $B^* = R_1^{-1} G$ หรือโดยอาศัย Cramer's Rule หรือโดยวิธีอื่น ๆ เช่นวิธีของ Gauss และ Gauss - Jordan

สำหรับวิธีที่จะเสนอในที่นี้จะใช้วิธี Cramer's Rule[1] เพราะสามารถนำไปสู่การคำนวณหา Partial Correlation และ Multiple Correlation (R^2) ได้โดยง่าย[2]

ให้ R คือ Correlation Matrix โดยที่

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ r_{31} & r_{32} & \cdots & r_{3k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{bmatrix}$$

---

[1] Cramer's Rule เป็นวิธีแก้สมการซึ่งเสนอโดยนักคณิตศาสตร์ชาวสวิสชื่อ Gabriel Cramer (1704-1752)

[2] ถ้าแก้โดยวิธีใช้ Matrix Inversion จะพบว่า จาก $R_1 B^* = G$ โดยที่ $G = s_1 R_{yj}$ เมื่อ $R_{yj} = (r_{12}, r_{13}, \ldots r_{1k})'$ โดยที่ r_{ij} คือสหสัมพันธ์ระหว่าง Y กับ X_j

ดังนั้น $B^* = R_1^{-1} G = s_1 R_1^{-1} R_{yj} = s_1 B^{**}$ เมื่อ $B^{**} = R_1^{-1} R_{yj}$

นั่นคือ $\hat{\beta}_j s_j = s_1 Bj^{**}$; $j = 2, 3, \ldots, k$ หรือ $\hat{\beta}_j = \frac{s_1}{s_j} Bj^{**}$; $j = 2, 3, \ldots, k$

โดยที่ $r_{ij}, i, j = 2, 3, ..., k$ คือสหสัมพันธ์ระหว่างตัวแปรสุ่ม X_i และ X_j และ $r_{j1} = r_{1j}$, $j = 2, 3, ..., k$ คือสหสัมพันธ์ระหว่างตัวแปรสุ่ม Y และ X_j และ $r_{ij} = 1$ เมื่อ $i = j = 1, 2, ..., k$

เมื่อพิจารณาระบบสมการที่ (3) จะพบว่า

$$B_2 = \frac{1}{s_2} \begin{vmatrix} r_{12}s_1 & r_{23} & \cdots & r_{2k} \\ r_{13}s_1 & r_{33} & \cdots & r_{3k} \\ \vdots & & & \\ r_{1k}s_1 & r_{3k} & \cdots & r_{kk} \end{vmatrix} \Big/ \begin{vmatrix} r_{22} & r_{23} & \cdots & r_{2k} \\ r_{23} & r_{33} & \cdots & r_{3k} \\ \vdots & & & \\ r_{2k} & r_{3k} & \cdots & r_{kk} \end{vmatrix} = \frac{s_1}{s_2} \begin{vmatrix} r_{12} & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{33} & \cdots & r_{3k} \\ \vdots & & & \\ r_{1k} & r_{3k} & \cdots & r_{kk} \end{vmatrix} \Big/ \begin{vmatrix} r_{22} & r_{23} & \cdots & r_{2k} \\ r_{23} & r_{33} & \cdots & r_{3k} \\ \vdots & & & \\ r_{2k} & r_{3k} & \cdots & r_{kk} \end{vmatrix}$$

[1]

$$= -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}$$ [2]

เมื่อ R_{12} คือ Cofactor ของ r_{12} ในเมตริกซ์ R และ R_{11} คือ Cofactor ของ r_{11} ในเมตริกซ์ R

---

[1] ถ้าเมตริกซ์ B เกิดขึ้นจากเมตริกซ์ A ที่สมาชิกทุกตัวในแถว (สดมภ์) ที่ i ของ A มีตัวคงที่ k เป็นตัวคูณร่วมแล้ว $|B| = k|A|$

[2] เมื่อ A เป็น Square Matrix ใด ๆ cofactor ของ a_{ij} คือ $A_{ij} = (-1)^{i+j} M_{ij}$ โดยที่ M_{ij} คือดีเทอร์มิแนนท์ของ Submatrix (เรียกว่า Minor Determinant) ที่เกิดขึ้นจากการตัดสมาชิกทุกตัวในแถวที่ i และสดมภ์ที่ j ของ A ทิ้ง ดังนั้น R_{1j} ก็คือ Cofactor ของ r_{1j} โดยที่ $R_{1j} = (-1)^{1+j}M_{1j}$ เมื่อ M_{1j} คือดีเทอร์มิแนนท์ของ Submatrix ที่เกิดขึ้นจากการตัดสมาชิกทุกตัวในแถวที่ 1 และสดมภ์ j ของ R ทิ้ง

และพบว่า $\hat{\beta}_j = \frac{1}{s_j} \frac{\begin{vmatrix} r_{22} & r_{23} & \cdots & r_{12}s_1 & \cdots & r_{2k} \\ r_{23} & r_{33} & \cdots & r_{13}s_1 & \cdots & r_{2k} \\ \vdots & \vdots & & & & \\ r_{2k} & r_{3k} & \cdots & r_{1k}s_1 & \cdots & r_{kk} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & \cdots & r_{2j} & \cdots & r_{2k} \\ r_{23} & r_{33} & \cdots & r_{3j} & \cdots & r_{3k} \\ \vdots & \vdots & & \vdots & & \vdots \\ r_{2k} & r_{3k} & \cdots & r_{kj} & \cdots & r_{kk} \end{vmatrix}} = -\frac{s_1}{s_j}\frac{R_{1j}}{R_{11}}$; $j = 2, 3, \ldots, k$[1]

โดยที่ R_{1j} ; $j = 2, 3, \ldots k$ คือ Cofactor ของ r_{1j} ในเมตริกซ์ R

ดังนั้น E(Y) หรือ Y สามารถแสดงได้อีกรูปหนึ่งดังนี้คือ

$$\hat{y} = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}x_2 - \frac{s_1}{s_3}\frac{R_{13}}{R_{11}}x_3 - \ldots - \frac{s_1}{s_k}\frac{R_{1k}}{R_{11}}x_k$$

หรือ
$$\frac{R_{11}}{s_1}\hat{y} + \frac{R_{12}}{s_2}x_2 + \frac{R_{13}}{s_3}x_3 + \ldots + \frac{R_{11}}{s_k}x_k = 0$$

ซึ่งในกรณีนี้จะพบว่า Residual Sum Square คือ $\sum_i^n e_i^2$ หรือ $e'e$ สามารถเสนอได้ในรูปของ R_{1j} ดังนี้

$$\sum_i^n e_i^2 = \sum_i^n (y_i - \hat{y}_i)^2$$

$$= \sum_i^n \{y_i - (-\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}x_{2i} - \frac{s_1}{s_3}\frac{R_{13}}{R_{11}}x_{3i} - \ldots - \frac{s_1}{s_k}\frac{R_{1k}}{R_{11}}x_{ki})\}^2$$

$$= \sum_i^n (y_i + \frac{s_1}{s_2}\frac{R_{12}}{R_{11}}x_{2i} + \frac{s_1}{s_3}\frac{R_{13}}{R_{11}}x_{3i} + \ldots + \frac{s_1}{s_k}\frac{R_{1k}}{R_{11}}x_{ki})^2$$

$$= \frac{1}{R_{11}^2}\sum_i^n (R_{11}y_i + \frac{s_1}{s_2}R_{12}x_{2i} + \frac{s_1}{s_3}R_{13}x_{3i} + \ldots + \frac{s_1}{s_k}R_{1k}x_{ki})^2$$

---

[1] R_{1j} คือ Cofactor ของ r_{1j} ดังนั้น $R_{1j} = (-1)^{1+j}M_{1j}$; $j = 2, 3, \ldots, k$

ดังนั้น $\hat{\beta}_j = \frac{-s_1}{s_j}\frac{R_{1j}}{R_{11}} = \frac{s_1}{s_j}(-1)^{1+j}M_{1j}/R_{11}$; $j = 2, 3, \ldots, k$

อ่านวิธีพิสูจน์ใน Hemmerle, William J., **Statistical Computation on A Digital Computer**, (Massachusetts, Blaisdell Publishing Company, 1967), p. 89-91 ซึ่งเสนอการพิสูจน์ในแนวที่แตกต่างกันกับวิธีที่เสนอไว้ในที่นี้แต่ผลลัพธ์ตรงกัน

$$= \frac{1}{R_{11}^2}\{(R_{11}^2 \sum_i^n y_i^2 + \frac{s_1^2}{s_2^2} R^2_{12} \sum_i^n x_{2i}^2 + \frac{s_1^2}{s_3^2} R_{13}^2 \sum_i^n x_{3i}^2 + \ldots + \frac{s_1^2}{s_k^2} R_{1k}^2 \sum_i^n x_{ki}^2)$$

$$+ (2\frac{s_1}{s_2} R_{11}R_{12} \sum_i^n x_{2i}y_i + 2\frac{s_1}{s_3} R_{11}R_{13} \sum_i^n x_{3i}y_i + \ldots + 2\frac{s_1}{s_{k-1}} \frac{s_1}{s_k} R_{1,k-1} R_{1k} \sum_i^n x_{ki}x_{(k-1)i})\}$$

$$= \frac{1}{R_{11}^2}[\{(n-1) R_{11}^2 s_1^2 + (n-1) R_{12}^2 s_1^2 + (n-1) R_{13}^2 s_1^2 + \ldots + (n-1) R^2{}_{1k} s_1^2\}$$

$$+ \{2(n-1) s_1^2 R_{11}R_{12}r_{12} + 2(n-1) s_1^2 R_{11}R_{13}r_{13} + \ldots + 2(n-1) s_1^2 R_{1,\,k-1}R_{1k}r_{k,\,k-1})\}]$$

$$= \frac{(n-1)\, s_1^2}{R_{11}^2}\{(R_{11}^2 + R_{12}^2 + \ldots + R_{1k}^2) + (2r_{12}R_{11}R_{12} + 2r_{13}R_{11}R_{13} + \ldots + 2r_{1k}R_{11}R_{1k} + \ldots$$

$$+ 2r_{k,\,k-1}R_{1,\,k-1}R_{1k})\}$$

$$= (n-1)\frac{s_1^2}{R^2{}_{11}}\{R_{11} \sum_{j=1}^k r_{1j}R_{1j} + R_{12} \sum_{j \neq s}^k r_{ij}R_{1s} + R_{13} \sum_{j \neq s}^k r_{ij}R_{1s} + \ldots + R_{1k} \sum_{j \neq s}^k r_{ij}R_{1s}\}$$

$$= \frac{(n-1)\, s_1^2}{R_{11}^2}\{R_{11}|R| + 0 + 0 + \ldots + 0\}$$

$$= \frac{(n-1)\, s_1^2\, |R|}{R_{11}}$$

หมายเหตุ โดยอาศัยทฤษฎี "ถ้ากระจายดีเทอร์มิแนนท์ตามสมาชิกของแถว (หรือสดมภ์) หนึ่งของเมตริกซ์ R แต่ใช้ Cafactor ในอีกแถว (หรือสดมภ์) หนึ่งแล้ว $|R| = 0$" โดยนัยนี้ $\Sigma\, r_{ij}R_{1s} = 0$

สำหรับความจริงตามทฤษฎีนี้ ถ้านักศึกษาไม่แน่ใจในความถูกต้องหรือเข้าใจไม่กระจ่างในวิธีพิสูจน์ให้นักศึกษาทดลองหา $\sum_i^n e_i^2$ กับเมตริกซ์ R ขนาดเล็กเช่นขนาด 3×3 หรือ 4×4 จะพบความจริงตามต้องการ

ดังนั้น Residual Sum Square สามารถเสนอในรูปของสหสัมพันธ์ได้ดังนี้คือ

$\sum_i^n e_i^2 = \frac{(n-1)\, s_1^2\, |R|}{R_{11}}$ เมื่อ R คือ Correlation Matrix และ R_{11} คือ Cofactor ของ r_{11}

$R_{11} = |R_1|$ กล่าวคือ

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \cdots r_{1K} \\ r_{21} & r_{22} & r_{23} \cdots r_{2K} \\ r_{31} & r_{32} & r_{33} \cdots r_{3K} \\ \vdots & \vdots & \vdots \quad \vdots \\ r_{K1} & r_{K2} & r_{K3} \cdots r_{KK} \end{bmatrix} \text{ โดยที่ } r_{jj} = 1 ; j = 1, 2, \ldots, k$$

และ $R_1 = \begin{bmatrix} r_{22} & r_{23} & \cdots & r_{2K} \\ r_{32} & r_{33} & \cdots & r_{3K} \\ \vdots & \vdots & & \vdots \\ r_{K2} & r_{K3} & \cdots & r_{KK} \end{bmatrix}$, $R_{11} = |R_1|$

โดยที่ R_1 คือ Coefficient Matrix ของสมการ $R_1 B^* = G$ ขอให้สังเกตว่าเมตริกซ์ R_1 เกิดจากการตัดสหสัมพันธ์ระหว่าง Y กับ X_j ทิ้งคือตัดแถวและสดมภ์ที่ 1 ของเมตริกซ์ R ทิ้ง R_1 จึงเป็นเมตริกซ์ของสหพันธ์ระหว่างตัวแปรอิสระเท่านั้น

และเนื่องจาก $R^2_{1\cdot 23\ldots k}$ [1] $= 1 - \dfrac{\sum_i^n e_i^2}{\sum_i^n y_i^2}$ โดยที่ $\sum_i^n e_i^2 = \dfrac{(n-1)\, s_1^2\, |R|}{|R_1|}$ และ $\sum_i^n y_i^2 = (n-1)s_1^2$

$$\text{ดังนั้น } R^2_{1\cdot 23 \ldots k} = 1 - \frac{|R|}{|R_1|} = 1 - \frac{|R|}{R_{11}}$$

สำหรับ Partial Correlation ระหว่างตัวแปรสุ่ม Y กับตัวแปรอิสระ $X_j ; j = 2, 3, \ldots, k$ คือ $R_{1j\cdot 23,\ldots,j-1,j+1,\ldots,k}$ สามารถพัฒนาได้ดังนี้

จากความรู้ในตัวอย่าง 2.2 เราทราบว่าถ้ามีสมการถดถอย $y = \hat{\beta}x$ และ $x = \hat{\beta}' y$ เรายอมสามารถหาค่าสหสัมพันธ์และ Variance Ratio ได้ดังนี้คือ

$$r^2 = \hat{\beta}\hat{\beta}', r = \sqrt{\hat{\beta}\hat{\beta}'}, \frac{s_x^2}{s_y^2} = \frac{\hat{\beta}'}{\hat{\beta}}$$

ซึ่งความรู้ส่วนนี้เป็นความรู้สำหรับกรณี Simple Regression ถ้าขยายแบบจำลองออกไปสู่กรณี

---

[1] ตอนที่ผ่านมาเราใช้สัญญลักษณ์ R^2 ในความหมายของ (Multiple Correlation)^2 ในที่นี้ใช้ $R^2_{1\cdot 23\ldots K}$ เพราะไม่ต้องการให้สับสนกับ R ที่หมายถึง Correlation Matrix ความจริงแล้วรูปที่สมบูรณ์ของ Multiple Correlation คือ $R_{1\cdot 23\ldots K}$ หรือ $\rho_{1\cdot 23\ldots K}$

ของ Multiple Regression และควบคุมอิทธิพลของตัวแปรที่ไม่เกี่ยวข้องให้คงที่เท่ากับค่าคงที่ใด ๆ เรายอมคำนวณหา (Coef. of Partial Correlation)[2] ได้โดยง่าย [1] ในที่นี้จะเริ่มจากกรณีที่มีตัวแปรอิสระ 2 ตัว 3 ตัว แล้วค่อยขยายสู่กรณีทั่วไปคือกรณีของตัวแปรอิสระ k - 1 ตัวดังนี้

1. กรณีสมการ $y = \beta_2 x_2 + \beta_3 x_3 + (u - \bar{u})$ จะพบว่า

(1) ถ้าต้องการหา $R^2_{12\cdot 3}$ ให้ควบคุมอิทธิพลของ x_3 ให้คงที่ซึ่งถ้าควบคุมให้ x_3 คงที่ไว้เท่ากับ c ค่าของ x_3 จะไปรวมกับค่าของ Y (ตามหลักเกณฑ์ของ Restricted Least Square) การวิเคราะห์สหสัมพันธ์ครั้งนี้จึงเปรียบเสมือนการวิเคราะห์ Simple Regression ที่มีตัวแปรอิสระเพียงตัวเดียว

ในขณะเดียวกันถ้าเราให้ x_2 ทำหน้าที่เป็นตัวแปรตามคือ $x_2 = \beta_2' y + \beta_3' x_3$ ซึ่งถ้าเราควบคุมให้ x_3 มีอิทธิพลคงที่คือมีค่าเท่ากับ c ค่าของ x_3 ก็จะไปรวมกับค่าของ x_2 ทำให้เราวิเคราะห์สมการถดถอยแบบ Simple Regression ที่มี Y เป็นตัวแปรอิสระ

ดังนั้นจากสมการ $y = \beta_2 x_2 + \beta_3 x_3 + (u - \bar{u})$ จะพบว่า

$$\hat{\beta}_2 = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}} \text{ และ } \hat{\beta}_3 = -\frac{s_1}{s_3}\frac{R_{13}}{R_{11}}$$

นั่นคือ
$$\hat{y} = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}} x_2 - \frac{s_1}{s_3}\frac{R_{13}}{R_{11}} x_3 \qquad \ldots\ldots\ldots (1)$$

และจากสมการ $x_2 = \beta_2' y + \beta_3' x_3 + (u - \bar{u})$ จะพบว่า

$$\hat{\beta}_2' = -\frac{s_2}{s_1}\frac{R_{21}}{R_{22}} \text{ และ } \hat{\beta}_3' = -\frac{s_3}{s_2}\frac{R_{23}}{R_{22}}$$

---

[1] Partial Correlation คือสหสัมพันธ์ระหว่างตัวแปรคู่ใด ๆ ที่คำนวณได้ภายใต้สถานการณ์ที่เราได้ควบคุมพฤติกรรม หรือความเคลื่อนไหวของตัวแปรอื่น ๆ ไว้ให้คงที่ นั่นคือถ้ามีตัวแปรอยู่ m ตัว และเราต้องการหาสหสัมพันธ์แท้ ๆ ของตัวแปรคู่ใด ๆ เราจะต้องควบคุมค่าของตัวแปรที่เหลือ m-2 ตัวให้คงที่ การควบคุมค่าของตัวแปรอื่น ๆ m-2 ตัวให้คงที่มีผลให้ค่าของตัวแปรเหล่านี้ถูกย้ายไปรวมกับค่าของตัวแปรตาม (ตามหลักของ Restricted Least Square) และสมการที่เหลืออยู่จึงเป็น Simple Regression เราจึงหาค่าสหสัมพันธ์ได้โดยใช้สูตร $r = \sqrt{\hat{\beta}_i \hat{\beta}_i'}$ ดังกล่าว แต่เนื่องจากสมการ Simple Regression นี้เกิดขึ้นจากการควบคุมความเคลื่อนไหวของตัวแปรอิสระอื่น ๆ เอาไว้ ค่าของ r ที่ใช้จึงมิใช่ Simple Correlation แต่เป็น Partial Correlation การหาสหสัมพันธ์คือ $R^2_{i,\cdot 23\ldots,j-1,j+1,\ldots,k}$ จึงให้เลือกเอา $\hat{\beta}_i$ และ $\hat{\beta}_i'$ มาคูณกันโดย $\hat{\beta}_i$ และ $\hat{\beta}_i'$ คือสัมประสิทธิความถดถอยของตัวแปรที่ไม่ได้ควบคุม

นั่นคือ $\hat{x}_2 = -\frac{s_2}{s_1}\frac{R_{21}}{R_{22}}y - \frac{s_3}{s_2}\frac{R_{23}}{R_{22}}x_3$ ..... (2)

ดังนั้น $R^2_{12\cdot 3} = \hat{\beta}_2\hat{\beta}_2' = \left(-\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}\right)\left(-\frac{s_2}{s_1}\frac{R_{21}}{R_{22}}\right) = \frac{R^2_{12}}{R_{11}R_{22}}$

(2) ถ้าต้องการหา $R^2_{13\cdot 2}$ ให้ควบคุมค่าของ x_2 ในสมการ $y = \beta_2 x_2 + \beta_3 x_3 + (u-\bar{u})$ และในสมการ $x_3 = \beta_2' x_2 + \beta_3' y + (u-\bar{u})$

ในสมการ $y = \beta_2 x_2 + \beta_3 x_3 + (u-\bar{u})$ พบว่า $\hat{\beta}_2 = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}$, $\hat{\beta}_3 = -\frac{s_1}{s_3}\frac{R_{13}}{R_{11}}$

ในสมการ $x_3 = \beta_2' x_2 + \beta_3' y + (u-\bar{u})$ พบว่า $\hat{\beta}_2' = -\frac{s_3}{s_2}\frac{R_{23}}{R_{33}}$, $\hat{\beta}_3' = -\frac{s_3}{s_1}\frac{R_{12}}{R_{33}}$

$$\therefore R^2_{13\cdot 2} = \hat{\beta}_3\hat{\beta}_3' = \left(-\frac{s_1}{s_3}\frac{R_{13}}{R_{11}}\right)\left(-\frac{s_3}{s_1}\frac{R_{13}}{R_{33}}\right)$$

$$= \frac{R^2_{13}}{R_{11}R_{33}}$$

2. กรณีสมการ $y = \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + (u-\bar{u})$ จะพบว่า

(1) ถ้าต้องการหา $R^2_{12\cdot 34}$ ให้ควบคุมค่าของ x_3, x_4 และพบว่า

ในสมการ $y = \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + (u-\bar{u})$ จะได้

$$\hat{y} = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}x_2 - \frac{s_1}{s_3}\frac{R_{13}}{R_{11}}x_3 - \frac{s_1}{s_4}\frac{R_{14}}{R_{11}}x_4$$

และในสมการ $x_2 = \beta_2' y + \beta_3' x_3 + \beta_4' x_4 + (u-\bar{u})$ จะได้

$$\hat{x}_2 = -\frac{s_2}{s_1}\frac{R_{12}}{R_{22}}y - \frac{s_2}{s_3}\frac{R_{23}}{R_{22}}x_3 - \frac{s_2}{s_4}\frac{R_{24}}{R_{22}}x_4$$

ดังนั้น $R^2_{12\cdot 34} = \left(-\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}\right)\left(-\frac{s_2}{s_1}\frac{R_{21}}{R_{22}}\right) = \frac{R^2_{12}}{R_{11}R_{22}}$

(2) ถ้าต้องการหา $R^2_{13\cdot 24}$ ให้ควบคุมค่าของ x_2 และ x_4 และพบว่า

ในสมการ $y = \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + (u-\bar{u})$ จะได้

$$\hat{y} = -\frac{s_1}{s_2}\frac{R_{12}}{R_{11}}x_2 - \frac{s_1}{s_3}\frac{R_{13}}{R_{11}}x_3 - \frac{s_1}{s_4}\frac{R_{14}}{R_{11}}x_4$$

หรือ

ในสมการ $x_3 = \beta_2' x_2 + \beta_3' y + \beta_4' x_4 + (u - \bar{u})$ จะได้

$$\hat{x}_3 = -\frac{s_3}{s_2}\frac{R_{23}}{R_{33}} x_2 - \frac{s_3}{s_1}\frac{R_{13}}{R_{33}} y - \frac{s_3}{s_4}\frac{R_{34}}{R_{33}} x_4$$

ดังนั้น $R^2_{13.24} = \left(-\frac{s_1}{s_3}\frac{R_{13}}{R_{11}}\right)\left(-\frac{s_3}{s_1}\frac{R_{13}}{r_{33}}\right) = \frac{R^2_{13}}{R_{11} R_{33}}$

(3) ถ้าต้องการหา $R^2_{14.23}$ ให้ควบคุมค่าของ x_2 และ x_3 และพบว่า

ในสมการ $x4 = \beta_2' x_2 + \beta_3' x_3 + \beta_4' y + (u - \bar{u})$

$$\hat{x}_4 = -\frac{s_4}{s_2}\frac{R_{24}}{R_{44}} x_2 - \frac{s_4}{s_3}\frac{R_{34}}{R_{44}} x_3 - \frac{s_4}{s_1}\frac{R_{14}}{R_{44}} y$$

ดังนั้น $R^2_{14.23} = \left(-\frac{s_1}{s_4}\frac{R_{14}}{R_{11}}\right)\left(-\frac{s_4}{s_1}\frac{R_{14}}{R_{44}}\right) = \frac{R^2_{14}}{R_{11} R_{44}}$

จากความรู้ตอนที่ 1 และ 2 ข้างต้นจะเห็นได้ว่า $R^2_{12.3} = \frac{R^2_{12}}{R_{11} R_{22}}$, $R^2_{13.2} = \frac{R^2_{13}}{R_{11} R_{33}}$,

$R^2_{12.34} = \frac{R^2_{12}}{R_{11} R_{22}}$, $R^2_{13.24} = \frac{R^2_{13}}{R_{11} R_{33}}$ และ $R^2_{14.23} = \frac{R^2_{14}}{R_{11} R_{44}}$

ด้วยเหตุนี้เราจึงสามารถขยายความมาสู่กรณีทั่วไปสำหรับสมการ

$y = \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + (u - \bar{u})$ ได้ดังนี้คือ

$$R^2_{1j.23..j-1,j+1,\ldots k} = \frac{R^2_{1j}}{R_{11} R_{jj}}; \quad j = 2, 3, \ldots, k$$

สำหรับเครื่องหมายของ $R_{1j.23\ldots j-1,\ j+1,\ \ldots k}$ จะมีเครื่องหมายเดียวกันกับ $\hat{\beta}_j$ และ $\hat{\beta}_j'$

ดังนั้น $R_{1j.23,\ldots j-1,j+1,\ldots k} = -4-j; \quad j = 2, 3, \ldots, k$

ตาราง ANOVA ซึ่งทดสอบโดยใช้ $R^2_{1.23\ldots k}$ เป็นเกณฑ์สามารถพัฒนาได้ดังนี้ ส่วนประโยชน์ของ Partial Correlation จะได้กล่าวถึงในลำดับต่อไป

จาก $\hat{\beta}' X' Y$ และ $e'e$ สำหรับกรณีของแบบจำลอง

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + (u_i - \bar{u}) \; ; \; i = 1, 2, \ldots, n$$

จะพบว่า $\dfrac{\hat{\beta}'X'Y}{Y'Y} = R^2_{1\cdot 23\ldots k}$

$$\hat{\beta}'X'Y = Y'Y \cdot R^2_{1\cdot 23\ldots k} = \sum_i^n y_i^2 R^2_{1\cdot 23\ldots k}$$

และ $\dfrac{e'e}{Y'Y} = R^2_{1\cdot 23\ldots k}$

$$e'e = Y'Y(1 - R^2_{1\cdot 23\ldots k}) = \sum_i^n y_i^2 (1 - R^2_{1\cdot 23\ldots k})$$

ตาราง ANOVA ปรากฏดังนี้

SOV	df	ss	MS	F
Regressor $X_2, X_3, \ldots, X_k$	k-1	$Y'Y \cdot R^2_{1\cdot 23\ldots k}$	$(Y'Y \cdot R^2_{1\cdot 23\ldots k})/k-1$	F_{c}
Residual	n-k	$Y'Y$	$(1-R^2_{1\cdot 23\ldots k})\{Y'Y(1-R^2_{1\cdot 23\ldots k})\}/n-k$	
Total	n - 1 .	$Y'Y$		

4.4.2 เทคนิคการวิเคราะห์แบบอื่น

จากสมการ $y = \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + (u - \bar{u})$ หรือ $Y = X\beta + (U - \bar{U})$ ซึ่งเราสามารถคำนวณหา Correlation Matrix R ได้ดังนี้คือ

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \ldots r_{1k} \\ r_{21} & r_{22} & r_{23} & \ldots r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \ldots r_{kk} \end{bmatrix}$$

เราสามารถใช้เมตริกซ์ R และเมตริกซ์ย่อยของ R ช่วยในการคำนวณเพื่อวิเคราะห์สมการถดถอย และวิเคราะห์สหสัมพันธ์ได้ดังนี้ [1]

[1] อ่านรายละเอียดใน Kerlinger,N.F. and Pedhazur, E.J., Ibid. p.53-99.

1. $\hat{\beta}_j = -\frac{s_1}{s_j}\frac{R_{1j}}{R_{11}}$; $j = 2, 3, \ldots, k$ เมื่อ R_{1j} คือ Cofactor ของ r_{1j} ในเมตริกซ์ R [1]

2. $R^2_{1.23\ldots k} = 1 - \frac{1}{r^{11}}$ เมื่อ r^{11} คือสมาชิก ณ ตำแหน่งที่ (1, 1) ของเมตริกซ์ R^{-1}

ค่า $R^2_{1.23\ldots k}$ ยังสามารถคำนวณได้จาก Semi Partial Correlation ได้อีก ซึ่งจะกล่าวต่อไป

3. $\hat{V}(\hat{\beta}_j) = \frac{\hat{\sigma}_2}{\sum x^2_j (1-R^2_j)}$; $j = 2, 3, \ldots, k$

โดยที่ $\hat{\sigma}_2 = \sum_i^n y_i^2 (1 - R^2_{1.23\ldots k})^2 / n-k$ และ $R^2_j = 1 - \frac{1}{r^{jj}}$ โดยที่ r^{jj} คือสมาชิก ณ ลำดับที่ $(j-1, j-1)$ ของเมตริกซ์ R_1^{-1} โดยที่ R_1 คือ Correlation matrix ระหว่างตัวแปรอิสระ $X_2, X_3, \ldots, X_k$

R^2_j เขียนให้เต็มรูปก็คือ $R^2_{j.\ 23\ldots j-1, j+1\ldots k}$ แสดงว่า R^2_j ก็คือกำลังสองของ Multiple Correlation ระหว่างตัวแปรอิสระ X_j กับตัวแปรอิสระอื่น ๆ ที่เหลืออยู่ การหาค่า R^2_j จึงกระทำเช่นเดียวกับการหา $R^2_{1\cdot 23\ldots k}$ ที่กล่าวแล้วข้างต้น R_1 คือ Correlation Matrix ที่เกิดขึ้นจากการตัดแถวที่ 1 และสดมภ์ที่ 1 ของเมตริกซ์ R ทิ้งกล่าวคือ

$$R_1 = \begin{bmatrix} r_{22} & r_{23} & r_{24} & \cdots & r_{2k} \\ r_{32} & r_{33} & r_{34} & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{k2} & r_{k3} & r_{k4} & \cdots & r_{kk} \end{bmatrix}$$

4. การทดสอบสมมุติฐาน

ก. $H_0 : \beta_2 = \beta_3 = \ldots\ldots = \beta_k = 0$ VS $H_1 : \beta_j$ ไม่เท่ากับศูนย์ทั้งหมด เราปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ $F_c > F_{1-\alpha, k-1, n-k}$ [2] โดยที่

$$F_c = \frac{R^2_{1.23\ldots k} / k-1}{(1-R^2_{1.23\ldots k})/n-k}$$

ข. $H_0 : \beta_j = 0$ VS $H_1 : \beta_j \neq 0$; $j = 2, 3, \ldots, k$ เราจะปฏิเสธสมมุติฐานหลัก ณ ระดับ

---

[1] หรือคำนวณหา $\hat{\beta}_j$ ได้จากเมตริกซ์ R^{-1} โดยที่ $\hat{\beta}_j = -\frac{r^{1j}}{r^{11}}$; $j = 2, 3, \ldots, k$ เมื่อ r^{ij} คือสมาชิก ณ ตำแหน่งที่ (i, j) ของเมตริกซ์ R^{-1} แต่วิธีนี้มี Rounding Error ค่อนข้างสูงจึงไม่ขอแนะนำให้ใช้

[2] โดยปกติ k คือจำนวนพารามิเตอร์ในสมการ $Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u$ ดังนั้น $k-1$ จึงหมายถึงเฉพาะจำนวนตัวแปรอิสระ X's หรือจำนวน Regression Coef. β's (ไม่นับรวม β_1) แต่ถ้าสมการถดถอยคือ $y = \sum_{j=2}^{k} \beta_j x_j + (u - \bar{u})$ $k-1$ จะหมายถึงทั้งจำนวนตัวแปรอิสระและจำนวนพารามิเตอร์ β

นัยสำคัญ α เมื่อ $|t_c| > t_{n-k, 1-\alpha/2}$ โดยที่

$$t_c = \hat{\beta}_j / \sqrt{\hat{V}(\hat{\beta}_j)} \; ; \; j = 2, 3, \ldots, k$$

ค. การตรวจสอบนัยสำคัญของตัวแปรที่นำเข้าสู่สมการ

โดยปกติการตรวจสอบนัยสำคัญของตัวแปรที่นำเข้าสู่สมการนั้นเรามีจุดมุ่งหมายเพื่อจะวัดดูว่าตัวแปรที่นำเข้ามาใหม่นั้นมีส่วนในการช่วยอธิบายความเคลื่อนไหวของค่า Y ได้อย่างจริงจัง (มิใช่ด้วยเหตุบังเอิญ) หรือไม่ ? มากน้อยเพียงใด ?

อย่างไรก็ตามการตรวจสอบนัยสำคัญของตัวแปรที่นำเข้าสู่สมการมิได้ให้ประโยชน์มากนักในแง่ของ Multiple Regression เพราะใน Multiple Regression นั้นเรานำตัวแปรทุกตัวเข้าสู่แบบจำลองพร้อมกันหมดมิใช่นำเข้าทีละตัวเช่น ที่ปฏิบัติกันใน Forward Procedure หรือ Stepwise Procedure เว้นแต่ผู้วิจัยจะใช้วิจารณญาณประกอบในการจัดเรียงตัวแปรหรือมีข้อสนเทศจาก Related Liturature เป็นเครื่องชี้ลักษณะการเรียงลำดับตัวแปร การตรวจสอบนัยสำคัญตามนัยนี้จึงมีประโยชน์อย่างแท้จริง ขอให้ระลึกไว้เสมอว่าตัวแปรอิสระตัวที่อยู่ใกล้ Y มากกว่าจะมีอิทธิพลต่อ Y มากกว่าตัวแปรที่อยู่ห่าง Y ออกไป เช่นในสมการ $Y = \hat{\beta}_1 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_2 X_2$ แสดงว่า X_3 มีอิทธิพลต่อ Y มากที่สุด X_4 มีอิทธิพลในลำดับรองขณะที่ X_2 มีอิทธิพลน้อยที่สุด ถ้าผู้วิจัยวิเคราะห์โดยใช้สมการ $Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$ ซึ่งแม้จะให้ค่า R^2 เท่ากัน ($R^2_{1.234} = R^2_{1.243} = R^2_{1.324} = R^2_{1.342} = R^2_{1.423} = R^2_{1.432}$) แต่ความหมายของตัวแปรในแง่ของอิทธิพลที่มีต่อ Y จะต่างกัน

ดังนั้นสำหรับ $H_0 : \beta s = 0$ VS $H_1 : \beta_s \neq 0$; $s = 2, 3, \ldots, k$ เมื่อ β_s คือสัมประสิทธิของ X ตัวที่เข้าสู่แบบจำลองล่าสุด (เข้าสู่แบบจำลองเป็นตัวที่ s) เราจึงตัดสินใจปฏิเสธสมมุติฐานหลักเมื่อ $F_c > F_{1-\alpha, k_1 - k_2, n-k_1}$ โดยที่

$$F_c = \frac{(R^2_{1.23 \ldots k_1} - R^2_{1.23 \ldots k_2}) / k_1 - k_2}{(1 - R^2_{1.23 \ldots k_1}) / n - k_1}$$

เมื่อ k_1 = จำนวนตัวแปรอิสระที่ใช้ในการวิเคราะห์หา $R^2_{1.23\ldots,s}$ และ k_2 คือจำนวนตัวแปรอิสระที่ใช้ในการวิเคราะห์หา $R^2_{1.23\ldots s-1}$

ตัวอย่างเช่น ในสมการ $y = \hat{\beta}_1 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_2 x_2$ เราจะมีความจำเป็นต้องทดสอบเพียงว่า β_4 และ β_2 มีนัยสำคัญหรือไม่เท่านั้น (ไม่ต้องตรวจสอบ β_3) ดังนี้

$$Ho : \beta_4 = 0 \quad VS \; H_1 : \beta_4 \neq 0 \quad F_c = \frac{(R^2_{1.34} - R^2_{1.3}) / 2-1}{(1 - R^2_{1.34}) / n-2}$$

$$H_0 : \beta_2 = 0 \text{ VS } H_1 : \beta_2 \neq 0 \quad F_c = \frac{(R^2_{1\cdot342} - R^2_{1\cdot34}) / 3-2}{(1-R^2_{1\cdot342}) / n-3}$$

5. Partial Correlation

Partial Correlation ($R_{1j\cdot23 \ldots j-1, j+1, \ldots, k}$) หมายถึงสหสัมพันธ์ระหว่างเฉพาะตัวแปร Y กับ X_j ล้วน ๆ เพราะได้ควบคุมอิทธิพลร่วมอื่น ๆ ที่ X_s ; $s = 2, 3, \ldots, k$; $s \neq j$ ได้นำเข้ามา คำว่าการควบคุมอิทธิพลของ X_s ; $s = 2, 3, \ldots, k$; $s \neq j$ นั้นเราหมายถึงการควบคุมอิทธิพลของ X_s ทั้ง 2 ทิศทางคือทั้งในตัวแปร Y และในตัวแปร X_j ซึ่งแสดงว่าอิทธิพลของตัวแปรอื่น ๆ ถูกควบคุมหรือคัดออก (partial out) จากตัวแปร Y และตัวแปร X_j และเนื่องจาก $R^2_{1\cdot23 \ldots k}$ คือ Variance Ratio ซึ่งสามารถเสนอได้ 2 รูปคือ

$$R^2_{1\cdot23\ldots k} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \text{\% Explained Variathon}$$

$$\text{หรือ } R^2_{1\cdot23\ldots k} = 1 - \frac{\sum e_i^2}{\sum y_i^2} = 1 - \text{\% Unexplained Variation}^{1}$$

พิจารณาฟังก์ชัน $Y = f(X_2, X_3, \ldots, X_k)$ ถ้าหากเราต้องหา $R^2_{1j\cdot23 \ldots, j-1, j+1, \ldots k}$ เราจะควบคุมหรือดึงอิทธิพลของ $X_2, X_3, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ ออกจากตัวแปร Y และ X_j ได้โดยอาศัย Multiple Regression Technique ด้วยการวิเคราะห์สมการถดถอย 2 สมการคือ

(1) $Y = f(X_2, X_3, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$ แล้วหา Residual Vector e_1 (ขนาด nxl) Residual e_1 จะแสดงอิทธิพลเฉพาะตัวของ Y เท่านั้น[2]

(2) $X_j = f(X_2, X_3, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$ แล้วหา Residual Vector e_j (ขนาด n×l) Residual e_j จะแสดงอิทธิพลเฉพาะตัวของ X, เท่านั้น

จากเวคเตอร์ e_1 และ e_j เราย่อมหาค่าสหสัมพันธ์ (Simple Linear Correlation) ได้ค่าสหสัมพันธ์นี้ก็คือ Partial Correlation $R_{1j\cdot23 \ldots, j-1, j+1, \ldots k}$ นั่นเอง

ด้วยเหตุนี้เราจึงสามารถคำนวณหา Partial Correlation ได้อีกวิธีหนึ่งซึ่งง่ายกว่าวิธีที่

---

[1] % Unexplained Variation $= 1 - R^2_{1\cdot23 \ldots k} = \frac{\sum e_i^2}{\sum y_i^2}$

[2] ถือว่าอิทธิพลอื่น ๆ ที่ผสมกันอยู่ในตัวแปรสุ่ม u ไม่มีผลมากนัก

กล่าวมา แล้วดังนี้[1]

$$R^2_{1j\cdot 23\ldots j-1,\,j+1,\ldots,k} = \frac{(1-R^2_{1\cdot 23\ldots j-1,\,j+1,\ldots,k}) - (1-R^2_{1\cdot 23\ldots j-1,\,j,\,j+1,\ldots,k})}{1-R^2_{1\cdot 23\ldots j-1,\,j+1,\ldots,k}}$$

ซึ่งเมื่อแจกแจงรายละเอียดตามสูตรนี้ออกมาจะพบว่า

$$\begin{aligned} R^2_{1j\cdot 23\ldots,\,j-1,\,j+1,\ldots,\,k} &= \frac{(\Sigma e^2_{2i}/\Sigma y^2_i) - (\Sigma e^2_{1i}/\Sigma y^2_i)}{\Sigma e^2_{2i}/\Sigma y^2_i} \\ &= \frac{\Sigma e^2_{ji}/\Sigma y^2_i}{\Sigma e^2_{2i}/\Sigma y^2_i} \\ &= \frac{\Sigma e^2_{ji}}{\Sigma e^2_{2i}} \\ &= \frac{s^{2*}_{xj}}{s^{2*}_{y}} \end{aligned}$$

โดยที่ e_2 คือ Residual Vector จากสมการ $Y = f(X_2, X_3, \ldots, X_{j-1}, X_{j+1}, \ldots X_K)$ ขณะที่ e_1 คือ Residual Vector จากสมการ $Y = f(X_2, X_3, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_k)$ และเมื่อพิจารณา e_2 และ e_1 จะพบว่า e_2 ยังคงมีอิทธิพลของ X_j แฝงอยู่ในขณะที่ใน e_1 นั้นอิทธิพลของตัวแปรอิสระทุกตัวถูกคัดออกหมดแล้ว ดังนั้นผลต่างระหว่าง $\Sigma e^2_{2i}/\Sigma y^2_i$ กับ $\Sigma e^2_{1i}/\Sigma y^2_i$ จึงแสดงถึงอิทธิพลเฉพาะตัวของ X_j เท่านั้น

สูตร Partial Correlation นี้สามารถนำไปใช้ได้ง่ายกว่าสูตรที่ผ่านมา ผู้เขียนจะแสดงวิธีใช้ให้เห็นในตัวอย่างที่ 4.3 ต่อไปนี้

ตัวอย่าง 4.3 จากการศึกษาทางด้านจิตวิทยา นักวิจัยอยากทราบว่าทัศนคติของคนเราที่มีต่อบุคคลอื่นสัมพันธ์และได้รับอิทธิพลมาจากลักษณะของการวางอำนาจ (X_2) ความหัวรั้น (X_3) และความเจ้าระเบียบ (X_4) หรือไม่เพียงใด ผลการบันทึกข้อมูลด้วยวิธีการที่เหมาะสม

---

[1] Partial Correlation สามารถคำนวณได้อีกวิธีหนึ่งคือ $R^2_{1j\cdot 23\ldots,\,j-1,\,j+1,\ldots\,k} = \frac{t_j^2}{t_j^2+(n-k)}$ โดยที่ $t_j = \hat{\beta}_j/\sqrt{\hat{v}(\hat{\beta}_j)}$ ซึ่งเป็น Test Statistics สำหรับ $H_0: \beta_j = 0$ VS $H_1: \beta_j \neq 0$ อ่านรายละเอียดใน Maddala, G.S., Op. Cit, p.110

[2] s^{2*}_y = Variance ของ Y ภายหลังจากเมื่อได้ขจัดอิทธิพลของ $X_2, X_3, \ldots, X_{j-1}, X_{j+1}..X_K$ ออกแล้ว

s^{2*}_{xj} = Variance ของ X_j ภายหลังเมื่อได้ขจัดอิทธิพลของ $X_2, X_3, \ldots, X_{j-1}, X_{j+1}..X_K$ ออกแล้วดูตัวอย่าง 2.2

ปรากฏข้อมูลดังนี้

																				รวม	เฉลี่ย	SS
Y	2	1	1	1	5	4	7	6	7	8	3	3	6	6	10	9	6	6	9	10	110	5.5	770
X_2	2	2	1	1	3	4	5	5	7	6	4	3	6	6	8	9	10	9	4	4	99	4.95	625
X_3	5	4	5	3	6	4	6	4	3	3	3	6	9	8	9	6	4	5	9	8	110	5.5	690
X_4	1	2	4	4	5	6	3	3	7	7	8	9	5	4	5	5	7	8	8	7	108	5.4	676
																					รวม	เฉลี่ย	SS

Deviation Sum of Square , Cross Product , สหสัมพันธ์และส่วนเบี่ยงเบนมาตรฐานของตัวแปรปรากฏดังตารางต่อไปนี้ ข้อมูลใต้แนว Main Diagonal คือค่าสหสัมพันธ์

	y	x_2	x_3	x_4
y	165.00	100.50	63.00	43.00
x_2	.6735	134.95	15.50	39.40
x_3	.5320	.1447	85.00	2.00
x_4	.3475	.3521	.0225	92.80
s	2.9469	2.6651	2.1151	2.2100

จงวิเคราะห์สมการถดถอย $Y = f(X's)$

วิธีทำ

วิธีที่ 1 จากสมการ $y = \beta_2 x_2+\beta_3 x_3+\beta_4 x_4+(u-\bar{u})$ พบว่า

$$\hat{\beta} = (X'X)^{-1}X'Y \text{ และ } \hat{\sigma}^2 = \frac{1}{n-k}(Y'Y-\hat{\beta}'X'Y)$$

และจากตารางเราพบว่า

$$(X'X) = \begin{bmatrix} 134.95 & 15.50 & 39.40 \\ 15.50 & 85.0 & 2.00 \\ 39.40 & 2.00 & 92.80 \end{bmatrix} , X'Y = \begin{bmatrix} 100.50 \\ 63.00 \\ 43.00 \end{bmatrix} , Y'Y = 165.00$$

จะพบว่า $(X'X)^{-1} = \begin{bmatrix} 8.64338347\,E-03 & -1.49055608\,E-03 & -3.63758832\,E-03 \\ & 1.20277219\,E-02 & 3.73625708\,E-04 \\ & & 1.23122169\,E-02 \end{bmatrix}$

ดังนั้น $\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} .618 \\ .624 \\ .187 \end{bmatrix}$ และ $\hat{\beta}'X'Y = 109.5134$, $R^2_{1.234} = .6637$

$$\because \hat{V}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} = (3.468)(X'X)^{-1}$$

ดังนั้น $\hat{V}(\hat{\beta}) = \begin{bmatrix} 2.99752538\,E-02 & -5.16924849\,E-03 & -1.26151563\,E-02 \\ & 4.17121398\,E-02 & 1.29573395\,E-03 \\ & & 4.26987682\,E-02 \end{bmatrix}$

ดังนั้นสมการถดถอยที่ต้องการคือ

$$y = \underset{(0.173)}{0.618\,x_2} + \underset{(0.204)}{0.624\,x_3} + \underset{(0.207)}{0.187\,x_4} : R^2_{1.234} = .6637$$

ตาราง ANOVA สำหรับ $H: \beta_2 = \beta_3 = \beta_4 = 0$ VS $H_1: \beta_j$ ไม่เท่ากับ 0 ทั้งหมดปรากฏดังนี้

SOV	df	SS	MS	F
X_2, X_3, X_4	3	109.52	36.506	10.526**
Residual	16	55.48	3.468	
Total	19	165.00		

$F_{3,16,.95} = 3.24$, $F_{3,16,.99} = 5.29$

วิธีที่ 2 จากตารางพบว่า $s_1 = 2.9469$, $s_2 = 2.6651$, $s_3 = 2.1151$, $s_4 = 2.2100$

$$R = \begin{bmatrix} 1 & .6735 & .5320 & .3475 \\ .6735 & 1 & .1447 & .3521 \\ .5320 & .1447 & 1 & .0225 \\ .3475 & .3521 & .0225 & 1 \end{bmatrix}$$

$$R_1 = \begin{bmatrix} 1 & .1447 & .3521 \\ .1447 & 1 & .0225 \\ .3521 & .0225 & 1 \end{bmatrix}$$

ซึ่งมีค่าดีเทอร์มิแนนท์ดังนี้ $|R| = .288122$, $R_{11} = |R_1| = .856874$

$$R_{12} = (-1)^{1+2} \begin{vmatrix} .6735 & .1447 & .3521 \\ .5320 & 1 & .0225 \\ .3475 & .0225 & 1 \end{vmatrix} = -0.47917 , \quad R_{13} = (-1)^{1+3} \begin{vmatrix} .6735 & 1 & .3521 \\ .5320 & .1447 & .0225 \\ .3475 & .3521 & 1 \end{vmatrix} = -0.38381$$

$$R_{14} = (-1)^{1+4} \begin{vmatrix} .6735 & 1 & .1447 \\ .5320 & .1447 & 1 \\ .3475 & .3521 & .0225 \end{vmatrix} = -0.1204$$

เนื่องจาก $\hat{\beta}_j = -\frac{s_1}{s_j}\frac{R_{1j}}{R_{11}}$; $j = 2, 3, 4$ และ $R^2_{1\cdot 234} = 1 - \frac{|R|}{R_{11}}$

ดังนั้น $\hat{\beta}_2 = \{\frac{-2.9469}{2.6651}\}\{\frac{-0.47917}{0.856874}\} = 0.618$

$\hat{\beta}_3 = \{\frac{-2.9469}{2.1151}\}\{\frac{-0.383810}{0.856874}\} = 0.624$

$\hat{\beta}_4 = \{\frac{-2.9469}{2.2100}\}\{\frac{-0.1204121}{0.856874}\} = 0.187$

$R^2_{1\cdot 234} = 1 - \frac{|R|}{R_{11}} = 1 - \frac{.288122}{.856874} = .6637$

$\hat{\beta}' X' Y = Y' Y \cdot R^2_{1\cdot 234} = \sum_i^n y^2_i \cdot R^2_{1\cdot 234} = (165.00)(.6637) = 109.52$

$Y' Y - \hat{\beta}' X' Y = Y' Y (1 - R^2_{1\cdot 234}) = \sum_i^n y_i^2 (1 - R^2_{1\cdot 234}) = (165.00)(.3363) = 55.48$

SOV	df	SS	MS	F
X_2, X_3, X_4	3	109.52	36.506	10.526**
Residual	16	55.48	3.468	
Total	19	165.00		

สำหรับเมตริกซ์ที่น่าสนใจและจะมีผลต่อการศึกษาต่อไปคือ R^{-1} และ R_1^{-1} ตามตัวอย่างจะพบว่า

$$R^{-1} = \begin{bmatrix} 2.973998 & -1.66308 & -1.332116 & -0.417921 \\ & 2.0964482 & .585304 & -.173408 \\ & & 1.619033 & .220396 \\ & & & 1.201326 \end{bmatrix}$$

$$R_1^{-1} = \begin{bmatrix} 1.166442 & -.159624 & -.407113 \\ & 1.022351 & .0332068 \end{bmatrix}$$

หมายเหตุ $R, R_1, R^{-1}, R_1^{-1}; X'X$ เป็น Symmetric Matrix ซึ่งสมมาตรกันในแนว Main Diagonal ในที่นี้ จึงเสนอไว้เพียงครึ่งรูป

จากเมตริกซ์ R^{-1} และ R_1^{-1} เราสามารถคำนวณหา $R^2_{1\cdot234....k}$ และ $\hat{V}(\hat{\beta}_j)$ ได้ดังนี้[1]

$$\hat{V}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum x_{ji}^2 (1-R_j^2)} \; ; j = 2, 3, \ldots, k$$

โดยที่ R_j^2 = Multiple Correlation ระหว่างตัวแปรอิสระ X_j กับตัวแปรอิสระอื่น ๆ ที่เหลืออยู่

$R_j^2 = 1 - \frac{1}{r^{jj}}$ เมื่อ r^{jj} คือสมาชิกในแนว Main Diagonal ของเมตริกซ์ R_1^{-1} และ

$t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$; $j = 2, 3, \ldots, k$ สำหรับทดสอบ $H_0 : \beta_j = 0$ VS $H_1 : \beta_j \neq 0$

จากข้อมูลเราทราบว่า

$$\sum_i^{20} x^2_{2i} = 134.95 \; , \; \sum_1^{20} x^2_{3i} = 85.00, \; \sum_i^{20} x^2_{4i} = 98.20$$

และจากเมตริกซ์ R_1^{-1} พบว่า

$$r^{22} = 1.66442, \; r^{33} = 1.022351 \; , \; r^{44} = 1.142597$$

$$R^2_{2\cdot34} = 1 - 1/1.166442 = .14269$$

$$R^2_{3\cdot24} = 1 - 1/1.022351 = .02186$$

---

[1] อ่าน Kerlinger, Fred N. and Pedhazur, Elazar J., Op. Cit.. p.67

$$R^2_{4\cdot23} = 1 - 1/1.142597 = .1248$$

$$\hat{V}(\hat{\beta}_2) = \frac{3.468}{134.95(1-R^2_{2\cdot34})} = .0300\,,\ t_2 = 3.5704$$

$$\hat{V}(\hat{\beta}_3) = \frac{3.468}{85(1-R^2_{3\cdot24})} = .0417\,,\ t_3 = 3.0558$$

$$\hat{V}(\hat{\beta}_4) = \frac{3.468}{98.20(1-R^2_{4\cdot23})} = .0427\,,\ t_4 = .9071$$

สำหรับเมตริกซ์ R^{-1} เราสามารถใช้ประโยชน์ในการหา $R^2_{1\cdot23\ldots k}$ และ $\hat{\beta}_j$ ได้ดังนี้

$R^2_{1\cdot23\ldots k} = 1 - \frac{1}{r^{11}}$ เมื่อ r^{11} = สมาชิก ณ ตำแหน่ง (1, 1) ของเมตริกซ์ R^{-1}

จากเมตริกซ์ R^{-1} จะพบว่า

$$R^2_{1\cdot234} = 1 - \frac{1}{2.973998} = .6637$$

จะเห็นได้ว่าการคำนวณหา $R^2_{1\cdot234}$ โดยวิธีนี้เป็นวิธีที่ง่ายและแสดงให้เห็นถึงการนำ Correlation Matrix มาใช้ประโยชน์ได้อย่างเต็มที่ นอกเหนือจากการใช้ประโยชน์ของ Correlation Matrix R และ R_1 แล้วเรายังสามารถใช้เมตริกซ์ R ให้เป็นประโยชน์ด้วย การนำไปใช้หา Partial Correlation ได้ดังนี้

จากสูตร $R^2_{1j\cdot23\ldots j-1,\,j+1,\,\ldots,\,k} = \frac{(1-R^2_{1\cdot23\ldots,\,j-1,\,j+1,\,\ldots\,k}) - (1-R^2_{1\cdot23\ldots k})}{(1-R^2_{1\cdot23\ldots,\,j-1,\,j+1,\,\ldots\,k})}$

แสดงว่า $R^2_{1\cdot23\ldots,\,j-1,\,j+1,\ldots k}$ สามารถคำนวณได้จากเมตริกซ์ R_j^{-1} เมื่อ R_j คือเมตริกซ์ที่เกิดขึ้นจากการตัดสมาชิกในแถวที่ j และสดมภ์ที่ j ของเมตริกซ์ R ทิ้งไป และ

$R^2_{1\cdot23\ldots j-1,\,j+1,\,\ldots,\,k} = 1 - \frac{1}{r^{11}}$ เมื่อ r^{11} คือสมาชิกลำดับที่ (1, 1) ของเมตริกซ์ R_j^{-1}

ดังนั้นในกรณีของตัวอย่างนี้เราจะพบว่า

$R^2_{12\cdot34} = \frac{(1-R^2_{1\cdot34}) - (1-R^2_{1\cdot234})}{1-R^2_{1\cdot34}}$ = Partial Correlation ระหว่าง Y กับ X_2 เมื่อขจัด

อิทธิพลของ X_3 และ X_4 ทิ้ง

$R^2_{13\cdot24} = \frac{(1-R^2_{1\cdot24}) - (1-R^2_{1\cdot234})}{(1-R^2_{1\cdot24})}$ = Partial Correlation ระหว่าง Y กับ X_3 เมื่อขจัด

อิทธิพลของ X_2 และ X_4 ทิ้ง

$$R^2_{14\cdot23} = \frac{(1-R^2_{1\cdot23}) - (1-R^2_{1\cdot234})}{(1-R^2_{1\cdot23})}$$ = Partial Correlation ระหว่าง Y กับ X_4 เมื่อขจัดอิทธิพลของ X_2 และ X_3 ทิ้ง

สำหรับการคำนวณหา Partial Correlation ระหว่างตัวแปรอิสระด้วยกันเองก็ให้กระทำในทำนองเดียวกันแต่ Correlation Matrix ที่จะนำมาใช้ในกรณีนี้คือเมตริกซ์ R_1 ซึ่งแสดงสหสัมพันธ์ระหว่างตัวแปรอิสระเท่านั้น

ความรู้ในเรื่องราวเหล่านี้จะมีประโยชน์โดยตรงต่อการวิเคราะห์ Multicolinearity ซึ่งจะได้กล่าวถึงในบทต่อไป

จากเมตริกซ์ $R = \begin{bmatrix} 1 & .6735 & .5320 & .3475 \\ & 1 & .1447 & .3521 \\ & & 1 & .1225 \\ & & & 1 \end{bmatrix}$ จะพบว่า[1]

$$R_4 = \begin{bmatrix} 1 & .6735 & .5320 \\ & 1 & .1447 \\ & & 1 \end{bmatrix}, \quad R_4^{-1} = \begin{bmatrix} 2.82861016143 & -1.723406236949 & -1.255443723395 \\ & 2.07141711728 & -.617118061208 \\ & & 1.578599077389 \end{bmatrix}$$

ดังนั้น $R^2_{1\cdot23} = 1 - \frac{1}{r^{11}} = 1 - 1/2.82861016143 = .6464$

$$R_3 = \begin{bmatrix} 1 & .6735 & .3475 \\ & 1 & .3521 \\ & & 1 \end{bmatrix}, \quad R_3^{-1} = \begin{bmatrix} 1.877953146795 & -1.181500824169 & -.2365822783211 \\ & 1.88485\ 1979166 & -.25308484546 \\ & & 1.171323515805 \end{bmatrix}$$

ดังนั้น $R^2_{1\cdot24} = 1 - \frac{1}{r^{11}} = 1 - 1/1.877953146795 = .4675$

$$R_2 = \begin{bmatrix} 1 & .5320 & .3475 \\ & 1 & .0225 \\ & & 1 \end{bmatrix}, \quad R_2^{-1} = \begin{bmatrix} 1.654701529793 & -.867802841452 & -.5554832176 \\ & 1.26880 & \\ & & \end{bmatrix}$$

---

[1] เนื่องจาก R เป็น Symmetric Matrix การเสนอ R จึงเสนอเพียงครึ่งรูป ส่วนที่ไม่ได้เสนอไว้จะสมมาตรกับตัวเลขที่ปรากฏอยู่เหนือแนว Main Diagonal

ดังนั้น $R^2_{1\cdot34} = 1 - \frac{1}{r^{11}} = 1 - \frac{1}{1.654701529793} = .3956$

นั่นคือ $R^2_{14\cdot23} = \frac{(1-R^2_{1\cdot23}) - (1-R^2_{1\cdot234})}{(1-R^2_{1\cdot23})}$

$\therefore R^2_{14\cdot23} = \frac{(1-.6464) - (1-.6637)}{1.6464} = .0489$; $R_{14\cdot23} = .2212$

$$R^2_{13\cdot24} = \frac{(1-R^2_{1\cdot24}) - (1-R^2_{1\cdot234})}{(1-R^2_{1\cdot24})}$$

$\therefore R^2_{13\cdot24} = \frac{(1-.4675) - (1-.6637)}{1-.4675} = .3685$, $R_{13\cdot24} = .6071$

$$R^2_{12\cdot34} = \frac{(1-R^2_{1\cdot34}) - (1-R^2_{1\cdot234})}{(1-R^2_{1\cdot34})}$$

$\therefore R^2_{12\cdot34} = \frac{(1-.3956) - (1-.6637)}{(1-.3956)} = .4436$, $R_{12\cdot34} = .6660$

หรือคำนวณหาค่า Partial Correlation ได้ด้วยสูตร

$$R^2_{1j\cdot23\ldots j-1, j+1, \ldots k} = \frac{t_j^2}{t_j^2+(n-k)} \quad ; j = 2, 3, \ldots, k \text{ ดังนี้}$$

จากสมการถดถอย $y = 0.618x_2 + 0.624x_3 + 0.187x_4$: $R^2 = .6637$

(0.173) (0.204) (0.207)

(3.572) (3.059) (0.903)

ดังนั้น $R^2_{12\cdot34} = \frac{t_2^2}{t_2^2+(n-k)} = \frac{(3.572)^2}{(3.572)^2+16} = .4436$

$R^2_{13\cdot24} = \frac{t_3^2}{t_3^2+(n-k)} = \frac{(3.059)^2}{(3.059)^2+16} = .3689$

$R^2_{14\cdot23} = \frac{t_4^2}{t_4^2+(n-k)} = \frac{(0.903)^2}{(0.903)^2+16} = .0485$

หมายเหตุ 1. ค่าในวงเล็บใต้สมการถดถอยแถวที่ 1 หมายถึง Standard Error ของ $\hat{\beta}_j$ แถวที่ 2 หมายถึงค่าของ t_j ซึ่ง $t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$; $j = 2, 3, \ldots, k$

2. ค่า Partial Correlation จาก 2 วิธีมีความคลาดเคลื่อนกันเล็กน้อยเนื่องมาจากการปัดเศษ (Rounding Error)

4.5 **การคัดเลือกสมการถดถอยที่ดีที่สุด** (Selecting the Best Regression Equation)

จากการศึกษาที่ผ่านมาจะพบว่าเราวิเคราะห์สมการถดถอยโดยนัยของ **Multiple Regression** มาโดยตลอดโดยถือว่าผู้วิจัยจะต้องใช้วิจารณญาณและความรู้ตัดสินว่าตัวแปรใดพึงนับได้ว่ามีความสำคัญต่อการกำหนดหรือควบคุมความเคลื่อนไหวของตัวแปรตาม (Y = Response)

อย่างไรก็ตาม เรายังไม่อาจแน่ใจได้ว่านักวิจัยทุกคนจะมีความรู้พอที่มองเห็นกลไกต่าง ๆ ที่ควบคุมความเคลื่อนไหวของ Y ได้เสมอไป ความเสี่ยงในการคัดเลือกหรือกำหนดตัวแปรอิสระที่มีลักษณะของ Linear Combination ของกันและกันคือ $X_j = \sum_{j=2}^{n} a_s X_s \; ; s \neq j$ จึงเกิดขึ้นได้และจะนำไปสู่ปัญหา Colinearity หรือ Multicolinearity ในที่สุด[1] นอกจากนี้ **Multiple Regression** ยังมีจุดอ่อนที่สำคัญอีก 2 ประการคือ

(1) ลำดับที่ของตัวแปรในแบบจำลองถูกจัดลำดับในเชิงอัตนิยม (Subjective Arrangement) ขึ้นอยู่กับความรู้ความสามารถของนักวิจัย อย่าลืมว่าตัวแปรอิสระที่ปรากฏในแบบจำลองนั้นเราถือว่าตัวแปรที่อยู่ในตำแหน่งที่ใกล้ตัวแปร Y มากกว่าจะมีอิทธิพลต่อ Y มากกว่าตัวแปรอิสระที่อยู่ห่างไกลออกไป เช่น

$$Y = \hat{\beta}_1 + \hat{\beta} X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

ในที่นี้เราจะถือว่า X_2 มีอิทธิพลต่อ Y มากที่สุด X_3 มีอิทธิพลในอันดับรองขณะที่ X_4 มีอิทธิพลต่อ Y น้อยที่สุด

ด้วยเหตุนี้ ถ้านักวิจัยจัดลำดับตัวแปรในเชิงอัตนิยม ความสับสนในลำดับความสำคัญของตัวแปรย่อมเกิดขึ้น สิ่งหนึ่งที่พึงเป็นที่สังเกตได้ก็คือ ถ้าเราจัดเรียงลำดับที่ของตัวแปรเสียใหม่ ซึ่งสามารถกระทำได้ถึง $(k-1)!$ วิธี (ตามตัวอย่างข้างต้นสามารถจัดลำดับได้ถึง $3! = 6$ วิธี) เราจะพบว่าตัวแปรบางตัว ซึ่งเคยมีนัยสำคัญจะกลับเป็นตัวแปรที่ไม่มีนัยสำคัญ (ค่า R^2 เท่ากันในทุกวิธี)[2]

---

[1] ปัญหา Multicolinearity คือปัญหาที่เกิดขึ้นเนื่องจากตัวแปรอิสระมีความสัมพันธ์ภายในต่อกันในอัตราค่อนข้างสูง หรือตัวแปรอิสระมีลักษณะของ Linear Combination ซึ่งมีผลให้ det $(X'X)$ มีค่าใกล้ศูนย์ (Near Singularity หรือ Singularity ถ้า det $(X'X) = 0$) อันจะมีผลสะท้อนให้ $\hat{\beta}$ มีคุณภาพต่ำลงหรือไม่อาจหาค่าในกรณีของ Singularity นักศึกษาจะได้ศึกษารายละเอียดในเรื่องนี้ต่อไป

[2] รูปเต็มของ R^2 คือ $R^2_{1.23\ldots k}$ ดังที่เสนอไว้ในตอน 4.4 ในตอน 4.4 จำเป็นเสนอรูปเต็มไว้เพราะเกรงว่านักศึกษาจะเข้าใจสับสนกับเมตริกซ์ R แต่โดยปกติเรานิยมใช้รูปย่อคือ R^2

(2) Multiple Regression สิ้นเปลืองค่าใช้จ่ายในการรวบรวมข้อมูล (Data) และข้อสนเทศ (Priori Information) สูงมาก ยิ่งนักวิจัยกำหนดตัวแปรอิสระไว้มากเพียงใด จำนวนข้อมูลและข้อสนเทศก็จะต้องเพิ่มขึ้นมากเพียงนั้นเพราะเราต้องกำหนดให้ $n >> k$ เพื่อควบคุมให้ Mean Square Residual มีค่าต่ำ แต่การเก็บรวบรวมข้อมูลโดยเฉพาะข้อมูลอนุกรมเวลามิใช่สิ่งที่นักวิจัยจะกำหนดจำนวนให้มากน้อยได้ตามความต้องการ ทั้งนี้เพราะจำนวนข้อมูลขึ้นอยู่กับแหล่งข้อมูลว่าจะมีข้อมูลให้อีกหรือไม่ แม้แต่ข้อมูลสำรวจหรือข้อมูลจากผลการทดลองก็เช่นกัน นักวิจัยจะเพิ่มขนาดตัวอย่างได้มากน้อยเพียงใดต้องขึ้นอยู่กับงบประมาณและเวลาเป็นสำคัญ นอกเหนือไปจากนี้แล้วการนำแบบจำลองที่ประกอบไปด้วยตัวแปรอิสระเป็นจำนวนค่อนข้างสูงไปใช้ในงานพยากรณ์ก็ยังประสบปัญหาและข้อจำกัดหลายประการโดยเฉพาะอย่างยิ่งก็คือข้อจำกัดเรื่องแบบแผนหรือค่าในอนาคตของ X's ซึ่งจะใช้เป็นฐานในการพยากรณ์ค่า Y ในอนาคตอีกด้วย

ด้วยเหตุนี้แบบจำลองที่ดีจึงควรเป็นแบบจำลองที่ใช้ตัวแปรอิสระไม่มากนัก แต่ทุกตัวต่างก็มีอิทธิพลต่อความเคลื่อนไหวของ Y อย่างแท้จริงสามารถแสดงลำดับและปริมาณของอิทธิพลที่มีต่อ Y ได้รวมทั้งให้ค่า R^2 สูงใกล้เคียงกับ Multiple Regression ด้วย

การศึกษาในลำดับต่อไปนี้ จะเป็นเรื่องของการคัดเลือก Subset ของตัวแปรอิสระที่สามารถใช้ควบคุมความเคลื่อนไหวของ Y ได้อย่างมีประสิทธิภาพ วิธีการเลือก Subset ดังกล่าวจำเป็นต้องอาศัยความรู้เรื่องสหสัมพันธ์ทั้ง Simple Correlation, Partial Correlation และ Multiple Correlation ซึ่งได้กล่าวถึงแล้วโดยละเอียดในตอนก่อน นอกจากนี้ยังต้องใช้ความรู้ประการอื่นผนวกเข้ามาอีกหลายประการเพื่อใช้เป็นเครื่องมือในการตัดสินว่าเราควรคัดเลือกตัวแปรใดไว้ในแบบจำลองและควรหยุดคัดเลือกตัวแปรเมื่อไร เช่น Partial F-Test, Mallaws' Cp Statistic, Tolerence, Colinearity และอื่น ๆ[1] ในที่นี้จะกล่าวถึงเฉพาะ Partial F-Test เท่านั้น

4.5.1 Extra Sum Square และ Partial F-Test

ให้ Z_j คือฟังก์ชันของตัวแปรอิสระ x, หรือ X_j's [2] และ

$$y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + ... + \beta_q Z_q + u \qquad \text{..... (1)}$$

คือสมการแสดงความสัมพันธ์ $Y = f(Z's)$

จากสมการ (1) เราสามารถหา OLS-Estimator ของ $_1\beta$ ได้ดังนี้

---

[1] อ่านรายละเอียดใน Weisberg, Sanford, Applied Linear Regression (John Wiley and Sons, NY, 1980), p.174-202

[2] ใช้ฟังก์ชันของ X เพื่อให้เกิดลักษณะของรูปทั่วไปตัวอย่างเช่น $Z_j = X_j^2, Z_j = e^{xj}$ etc.

${}_1\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q)$ ได้พร้อมทั้งสามารถหา Sum Square ที่เกี่ยวข้องได้ดังนี้คือ

(1) ${}_1\hat{\beta} = ({}_1Z'\,{}_1Z)^{-1}\,{}_1Z'Y$ เมื่อ ${}_1Z$ คือเมตริกซ์ขนาด n×q

(2) $SSR_1 = {}_1\hat{\beta}'\,{}_1Z'Y$

(3) $SSE_1 = Y'Y - {}_1\hat{\beta}'\,{}_1Z'Y$ และ $MSE_1 = \hat{\sigma}_1^2 = \frac{1}{n-q}(Y'Y - {}_1\hat{\beta}'\,{}_1Z'Y)$

ให้ $Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \ldots\ldots + \beta_q Z_q + \beta_{q+1} Z_{q+1} + \ldots\ldots + \beta_p Z_p + u$ ..... (2)

คือสมการแสดงความสัมพันธ์ $Y = f(Z's)$ โดยที่ $p > q$

จากสมการ (2) เราสามารถหา OLS-Estimator ของ ${}_2\beta$ คือ

${}_2\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q, \hat{\beta}_{q+1}, \ldots, \hat{\beta}_p)$ พร้อมทั้งสามารถคำนวณหา Sum Square ที่เกี่ยวข้องได้ดังนี้ คือ

(1) ${}_2\hat{\beta} = ({}_2Z_2'Z)^{-1}\,{}_2Z'Y$ เมื่อ ${}_2Z$ คือเมตริกซ์ขนาด n×p

(2) $SSR_2 = {}_2\hat{\beta}'\,{}_2Z'Y$

(3) $SSE_2 = Y'Y - {}_2\hat{\beta}'\,{}_2Z'Y$ และ $MSE_2 = \hat{\sigma}_2^2$ หรือ $s^2 = \frac{1}{n-p}(Y'Y - {}_2\hat{\beta}'\,{}_2Z'Y)$

จากผลลัพธ์ข้างต้นจะพบว่า Extra Sum Square Regression คือ

$ESSR = SSR_2 - SSR_1 = {}_2\hat{\beta}_2'Z'Y - {}_1\hat{\beta}_1'Z'Y$ ซึ่งเป็น Sum Square ของตัวแปรอิสระ (Regressor) $Z_{q+1}, Z_{q+2}, \ldots, Z_p$ ที่เพิ่มขึ้นมาจากสมการ(1)โดยที่สัมประสิทธิความถดถอยคือ $\hat{\beta}_{q+1}, \hat{\beta}_{q+2}, \ldots, \hat{\beta}_p$

โดยอาศัยความรู้เรื่อง Distribution of Quadratic Form เราสามารถพิสูจน์ได้ว่า[1]

$$\frac{SSR_1}{\sigma^2} \sim \chi^2_q \cdot \frac{SSR_2}{\sigma^2} \sim \chi^2_p \text{ และ } \frac{ESSR}{\sigma^2} \sim \chi^2_{p-q}$$

ดังนั้น $$\frac{ESSR/(p-q)\sigma^2}{(Y'Y - SSR_2)/(n-p)\sigma^2} = \frac{ESSR/(p-q)}{\hat{\sigma}_2^2} \sim F_{p-q,\,n-p}$$ [2]

ดังนั้นเราจึงปฏิเสธสมมุติฐาน

$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$ VS $H_1 : \beta_j$ ไม่เท่ากับ 0 ทั้งหมด; $j = q+1, q+2, \ldots, p$

ณ ระดับนัยสำคัญ α เมื่อ $F_c > F_{1-\alpha, p-q, n-p}$ โดยที่ $F_c = \dfrac{(SSR_2 - SSR_1)/p-q}{\hat{\sigma}_2^2}$

---

[1] อ่าน Graybill, F.A., Op. Cit.

[2] จากนิยาม $U = \dfrac{\chi^2_{1}/\nu_1}{\chi^2_{2}/\nu_2} \sim F_{\nu_1, \nu_2}$

จากความรู้ Extra Sum Square และการทดสอบสมมุติฐาน $H : \beta_{q+1} = \ldots = \beta_p = 0$ นี้นอกจากนำไปใช้สำหรับทดสอบหา Best Subset แล้วยังเป็นกรณีทั่วไปของ Partial F-Test และ Sequential F-Test ซึ่งใช้เป็นหลักในการพัฒนาวิธีการเลือกสมการถดถอยที่ดีที่สุดเช่น Backward Elimination, Forward Selection, Stepwise Regression และอื่น ๆ

1. Partial F-Test

จากสมการ $Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \ldots + \beta_p Z_p + u$ เราสามารถหา MSE และ Sum Square Regression ของเฉพาะ β_j ได้ดังนี้

$$MSE = \hat{\sigma}^2 = \frac{1}{n-p}(Y'Y - \hat{\beta}'Z'Y)$$

$$SS(\beta_j|\beta_1,\beta_2,\ldots\beta_{j-1},\beta_{j+1},\ldots\beta_p) = SS(\beta_1,\beta_2,\ldots\beta_p) - SS(\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p) ; 2 \leqslant j \leqslant p$$

หมายเหตุ

$SS(\beta_j|\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p)$ อ่านว่า Sum Square ของ X_j เมื่อ X's อื่น ๆ เข้าสู่แบบจำลองแล้ว

ดังนั้น Partial F คือ

$$F_c = \frac{SS(\beta_1,\beta_2,\ldots,\beta_p) - SS(\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1}\ldots,\beta_p)}{\hat{\sigma}^2}$$

โดยเราจะปฏิเสธ $H_0 : \beta j = 0$ Vs $H_1 : \beta_j \neq 0 ; 2 \leqslant j \leqslant p$ ณ ระดับนัยสำคัญ α เมื่อ

$$F_c > F_{1-\alpha,1,n-p}$$

Partial F – Test ใช้สำหรับตรวจสอบนัยสำคัญของ β_j เพื่อตัดสินใจว่าตัวแปรใดควรคงไว้ ตัวแปรใดควรตัดทิ้ง[1] โดยที่ β_j ปรากฏอยู่ ณ ตำแหน่งใดในแบบจำลองก็ได้ แต่ในทางปฏิบัติจะต้องยึดถือหลักเกณฑ์ของ Extra Sum Square ไว้เป็นแนวทางเสมอกล่าวคือ เราจะคำนวณหาค่า $SS(\beta_j \mid \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ ได้เฉพาะเมื่อจัดให้ X_j เข้าสู่สมการเป็นตัวสุดท้าย นั่นคือ $SS(\beta_j|\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ คำนวณหาได้จากผลต่างระหว่าง Sum Square Regression จากสมการ

[1] เหตุการณ์ที่เกิดขึ้นเป็นปกติก็คือตัวแปรที่เข้าสู่แบบจำลองแล้ว (บางตัวผ่านการคัดเลือกด้วย r_{ij}) มักจะกลับไม่มีนัยสำคัญเมื่อมีตัวแปรใหม่เข้าสู่สมการ การทดสอบว่าตัวแปรใดควรคงไว้หรือควรตัดทิ้งจึงมีความสำคัญต่อการวิเคราะห์มาก

$$Y=\beta_1+\beta_2Z_2+\ldots+\beta_{j-1}Z_{j-1}+\beta_{j+1}Z_{j+1}+\ldots+\beta_pZ_p+u$$

และ $$Y=\beta_1+\beta_2Z_2+\ldots+\beta_{j-1}Z_{j-1}+\beta_{j+1}Z_{j+1}+\ldots+\beta_pZ_p+\beta_jZ_j+u$$

2. Sequential F - Test

Sequential F - Test เป็นกรณีเฉพาะของ Partial F - Test ใช้เป็นเครื่องมือตรวจสอบนัยสำคัญของตัวแปรอิสระเข้าสู่สมการล่าสุดเท่านั้น กล่าวคือ

$$\text{Sequential F คือ } F_c=\frac{SS(\beta_1,\beta_2,\ldots,\beta_{p-1},\beta_p)-SS(\beta_1,\beta_2,\ldots,\beta_{p-1})}{\hat{\sigma}^2}$$

โดยที่ $\hat{\sigma}^2$ คือ MSE จากสมการ $Y=\beta_1+\beta_2Z_2+\ldots+\beta_pZ_p+u$ หรือ $\hat{\sigma}^2=\frac{1}{n-p}(Y'Y-\hat{\beta}'Z'Y)$ และจะปฏิเสธสมมุติฐาน $H_0:\beta_p=0$ VS $H_1:\beta_p\neq 0$ ณ ระดับนัยสำคัญ α เมื่อ

$$F_c>F_{1-\alpha,1,n-p}$$

4.5.2 เทคนิคการคัดเลือกตัวแปร

เทคนิคการคัดเลือกตัวแปร หรือเทคนิคการสร้างสมการถดถอยที่ดีที่สุด (Best Fitted Model) มีหลายวิธี ในที่นี้จะกล่าวถึงเฉพาะวิธี Forward Selection และ Stepwise Regression เท่านั้นเพราะเป็นวิธีที่พิสูจน์แล้วว่าเป็นวิธีที่ดีกว่าวิธีอื่น[1]

4.5.2.1 Forward Selection Procedure (FS)

การคัดเลือกตัวแปรโดยวิธี FS มีหลักการโดยสรุปคือนำตัวแปรอิสระเข้าสู่สมการคราวละ 1 ตัว โดยเริ่มจาก Simple Regression ก่อน แล้วค่อย ๆ เพิ่มตัวแปรอื่น ๆ เข้าสู่สมการคราวละตัวจนกว่าจะพบว่าตัวแปรอิสระอื่น ๆ ที่เหลืออยู่ไม่มีนัยสำคัญ หรือถ้านำเข้าสู่สมการแล้วจะก่อให้เกิดปัญหา Multicolinearity[2] จึงหยุดดำเนินการ

---

[1] อ่านรายละเอียดใน Drapper, N and Smith,H. Op. Cit , p.163-177 และ Weisberg , S., Op. cit , p.174-199

[2] Weisberg, S., Ibid., p.191 หลักการโดยสรุปของเรื่องนี้ก็คือให้วิเคราะห์สมการถดถอยระหว่างตัวแปรอิสระอื่น ๆ ที่เหลืออยู่กับตัวแปรที่เข้าสู่สมการแล้ว แล้ววัดดูว่า R^2 มีค่าสูงเกินไปหรือไม่ โดยปกติจะถือว่า ถ้า $R^2=.99$ หรือ .999 คือดัชนีที่ชี้ว่าเรากำลังเผชิญปัญหา Multicolinearity หรือจะใช้ Tolerence เป็นดัชนีก็ได้โดยที่ Tolerence $=1-R^2$ ถ้าพบว่า Tolerence มีค่าเท่ากับ .01 หรือ .001 แสดงว่าเรากำลังเผชิญปัญหา Multicolinearity วิธีนี้ไม่เป็นที่นิยมมากนัก เพราะมีลักษณะโน้มเอียงไปในเชิงอัตนิยม

1. เริ่มนำตัวแปรอิสระที่มีสหสัมพันธ์กับ Y สูงที่สุดเข้าสู่สมการ (ค่าสหสัมพันธ์ดูได้จากแถวที่ 1 หรือสดมภ์ที่ 1 ของเมตริกซ์ R) แล้ววัดค่า R^2 และตรวจสอบนัยสำคัญด้วย t-test หรือ F-test

2. เพิ่มตัวแปรอิสระอื่น ๆ เข้าสู่สมการในข้อ 1 คราวละ 1 ตัว ถ้าพบว่าตัวแปรดังกล่าวมีคุณสมบัติสอดคล้องกับเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้

(ก) มีค่า Partial Correlation กับตัวแปร Y สูงที่สุด ค่า Partial Correlationนี้คือสหสัมพันธ์ระหว่างตัวแปร Y กับตัวแปรอิสระภายนอก โดยที่เราได้ควบคุมอิทธิพลของตัวแปรอิสระที่เข้าสู่ สมการก่อนหน้านั้นเอาไว้

(ข) เพิ่มค่า R^2 สูงกว่าตัวแปรอิสระอื่น ๆ

(ค) มีค่า t หรือ Partial F [1] สูงกว่าตัวแปรอิสระอื่น ๆ โดยที่ t ต้องมีค่าสูงกว่าค่าวิกฤต และ Partial F ต้องมีค่าสูงกว่า $F_{1-\alpha,1,n-p}$หรือ F-IN

3. หยุดดำเนินการเมื่อพบว่าตัวแปรอิสระที่เหลืออยู่นอกสมการไม่มีนัยสำคัญทางสถิติ (ใช้ Partial F - Test) หรือให้ค่า Tolerence $\leq .01$ หรือ .001 หรือไม่ทำให้ R^2 มีค่าสูงขึ้น หรือสูงขึ้นไปเพียงเล็กน้อย

4.5.2.2 Stepwise Regression (SW)

SW เป็นวิธีปรับปรุงแบบจำลองวิธีหนึ่งที่ได้รับความนิยมใช้อย่างกว้างขวาง ความจริงแล้ว SW ก็คือ FS ที่ปรับปรุงแล้วนั่นเอง หรือถ้าจะมองอีกมุมหนึ่ง SW ก็คือวิธีที่ปรับปรุงแบบจำลองที่นำเอาทั้ง FS และ BE มาผสมผสานกัน กล่าวคือเมื่อตัวแปรอิสระได้รับคัดเลือก และนำเข้าสู่สมการแล้วตัวแปรอิสระทุกตัวในสมการจะต้องผ่านการพิจารณาเพื่อคัดออก ตัวแปรที่ถูกคัดออกถือว่าเป็นตัวแปรส่วนเกิน (Superfluous) เนื่องจากไม่มีนัยสำคัญ หรือเป็นตัวการที่จะนำไปสู่ปัญหา Multicolinearity ที่ต้องกระทำการตรวจสอบทั้งเพื่อ "คัดเข้า" และ "คัดออก" ก็เพราะลำดับที่ของตัวแปรอิสระในสมการมีผลต่อคุณค่าของตัวแปรและสมการดังที่กล่าวมาแล้วในตอนต้นความจริงที่ปรากฏอยู่เสมอในงานคัดเลือกตัวแปรก็คือตัวแปรตัวหนึ่งเคยมีอิทธิพลต่อ Y เป็นอย่างมากในขั้นหนึ่ง แต่ในภายหลังหลังจากที่นำตัวแปรอื่น ๆ เพิ่มเข้าสู่สมการกลับพบว่าตัวแปรนั้นกลับไม่มีนัยสำคัญ หรือกลับเป็นตัวแปรที่เกิดขึ้น (ที่มีความโน้มเอียง) จากการ

---

[1] ความจริงแล้วต้องเป็น Sequential F แต่เนื่องจาก Sequential F เป็นกรณีเฉพาะของ Partial F ในที่นี้จึงใช้ Partial F โดยตลอดเพื่อป้องกันความสับสน

ประกอบกันของตัวแปรอื่น ๆ (Linear Combination) ซึ่งในกรณีเช่นนี้ตัวแปรดังกล่าวต้องถูกตัดทิ้งเพราะถือว่าเป็นส่วนเกินหรือตัวก่อปัญหา[1]

หลักเกณฑ์ของ SW ปรากฏดังนี้

1. เริ่มนำตัวแปรอิสระที่มีสหสัมพันธ์กับตัวแปร Y สูงที่สุดเข้าสู่สมการแล้ววัดค่า R^2 และตรวจสอบนัยสำคัญด้วย t-test หรือ F-test[2]

2. เพิ่มตัวแปรอิสระอื่นเข้าสู่สมการในข้อ 1 คราวละ 1 ตัว ถ้าพบว่าตัวแปรดังกล่าว มีคุณสมบัติสอดคล้องกับเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้คือ

(ก) มีค่า Partial Correlation กับตัวแปร Y สูงที่สุด

(ข) เพิ่มค่า R^2 สูงกว่าตัวแปรอิสระอื่น ๆ

(ค) มีค่า t หรือ Partial F[3] สูงกว่าตัวแปรอิสระอื่นโดยที่ $t > t_{tab}$ และ Partial F > $F_{1-\alpha, 1, n-p}$ หรือ F-IN

3. ในทุกครั้งที่เพิ่มตัวแปรอิสระเข้าสู่สมการอันมีผลให้ในสมการมีจำนวนตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไปให้พิจารณาคัดตัวแปรอิสระที่อาจเป็นส่วนเกินทิ้งไป ตัวแปรที่จะต้องถูกคัดทิ้งคือตัวแปรที่มีคุณสมบัติสอดคล้องกับเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้ [4]

---

[1] เทคนิคการคัดเลือกตัวแปรที่มีลักษณะการดำเนินการสวนทางกับ FS คือ BE (Backward Elimination) ซึ่งมีวิธีการโดยสรุปคือ ให้นำตัวแปรอิสระทุกตัวเข้าสู่สมการโดยวิเคราะห์ตามวิธีของ Multiple Regression จากนั้นให้ค่อย ๆ พิจารณาคัดตัวแปรอิสระที่ไม่จำเป็นทิ้ง ตัวแปรใดมีคุณสมบัติสอดคล้องกับเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้จะถูกคัดทิ้ง

(ก) มี Partial Correlation กับตัวแปร Y น้อยที่สุด

(ข) ถ้าคัดออกแล้วจะมีผลให้ R^2 มีค่าลดลงไปน้อยกว่าตัวแปรอื่น

(ค) มีค่า t หรือ Partial F ต่ำกว่าตัวแปรอื่น ๆ หรือนัยหนึ่งคือไม่มีนัยสำคัญทางสถิติเมื่อตรวจสอบด้วย t-Test หรือ Partial F-Test โดยที่ $t < t_{tab}$ หรือ Partial F < $F_{1-\alpha, 1, n-p}$ หรือ F-OUT

[2] $F_{1,\nu} = t^2_{\nu}$ ในกรณีของขั้นที่ 1 นี้ไม่ควรใช้ Partial F-Test แต่ถ้าจะใช้ลักษณะของ F-IN และ F-OUT ซึ่งไม่ต่างจาก Partial F-Test นักก็เป็นสิ่งที่เป็นไปได้

[3] ค่า F ของตัวแปรที่เข้าสู่สมการเป็นตัวสุดท้าย (ล่าสุด) เรียกว่า Sequential F ความจริงแล้ว Sequential F ก็คือ Partial F นั่นเอง

[4] ขอให้สังเกตว่าขั้นที่ 1 และ 2 คือ FS ส่วนขั้นที่ 3 คือ BE

(1) มีค่า Partial Correlation กับตัวแปรอิสระ Y ต่ำที่สุด

(ข) มิได้ช่วยให้ R^2 ของสมการสูงขึ้น หรือถ้าคัดออกแล้วทำให้ค่า R^2 เดิมลดลงน้อยกว่าการคัดตัวแปรอิสระอื่น ๆ ทิ้ง (หมายความว่ามีความสำคัญน้อยที่สุด)

(ค) มีค่า t หรือ Partial F ต่ำกว่าตัวแปรอิสระอื่น ๆ โดยที่ $t < t_{tab}$ หรือ Partial $F < F_{1-\alpha, 1, n-p}$ หรือ F - OUT

4. หยุดดำเนินการเมื่อพบว่าตัวแปรอิสระภายนอกสมการมิได้มีนัยสำคัญทางสถิติ หรือมิได้เพิ่ม R^2

หมายเหตุ

1. โดยปกติการคัดตัวแปรอิสระเข้าสู่สมการหรือคัดตัวแปรอิสระออกจากสมการเรานิยมใช้ Partial F - Test มากกว่าเงื่อนไขอื่น ๆ เพราะสะดวกสบาย และไม่ต้องอาศัยวิจารณาญาณมากนัก แต่ด้วยเหตุที่ค่าวิกฤตของ F จะต้องเปลี่ยนแปลงไปตามจำนวน degree of freedom เสมอ กล่าวคือค่าวิกฤต $F_{1-\alpha,1, n-p}$ จะเปลี่ยนค่าไปตาม df ของ Denominator คือ n-p ซึ่งเปลี่ยนแปลงตามค่า p เมื่อ p คือจำนวนพารามิเตอร์ β ด้วยเหตุนี้เพื่อความสะดวกเราจึงนิยมใช้ค่าคงที่ คือ F-IN และ F - OUT แทนค่าวิกฤตของ F เพื่อผู้วิจัยจะได้ไม่ต้องเปิดตาราง F บ่อย ๆ เช่นกำหนดให้มีค่าระหว่าง 2-4

2. F-IN หรือเรียกอีกอย่างหนึ่งว่า F - Value for entry และ F-OUT หรือเรียกอีกอย่างหนึ่งว่า F - Value for deletion นั้นโดยปกติควรกำหนดให้มีค่าเท่ากัน[1] แต่ด้วยเหตุที่ F - IN และ F - OUT เป็นเงื่อนไขสำคัญที่ควบคุมกลไกของ SW การกำหนดค่าของ F - IN และ F - OUT จึงต้องกระทำอย่างรัดกุม กล่าวคือถ้าเรากำหนดให้ F - IN มีค่าต่ำมาก ๆ แบบจำลองก็จะรับตัวแปรอิสระเกือบทุกตัวเอาไว้ ถ้ากำหนดให้ F - OUT มีค่าสูงมาก ๆ ตัวแปรอิสระที่เข้าสู่สมการแล้วก็มีโอกาสปรากฏอยู่ในสมการโดยไม่ต้องถูกคัดออก ผลเสียที่เด่นชัดก็คือแบบจำลองอาจมีตัวแปรที่เป็นส่วนเกินปะปนอยู่เป็นจำนวนมาก จากการศึกษาของเคเนดี้และแบนครอฟท์ (Kenedy and Bancroft, 1971)[2] พบว่าพื้นที่ในเขตวิกฤตของ F ที่เหมาะสมคือ 25% สำหรับ F - IN และ 10% สำหรับ F - OUT การตัดสินใจให้กระทำดังนี้

(1) นำตัวแปรตัวใหม่เข้าสู่สมการถ้าตัวแปรนั้นให้ค่า Partial F (Sequential F) สูงกว่า F - IN

(2) คัดตัวแปรที่เข้าสู่สมการแล้วออกไปถ้าตัวแปรนั้นให้ค่า Partial F ต่ำกว่า F - OUT[3]

---

[1] Drapper, N.and Smith, H., Op. Cit., p.180

[2] Weisberg, S., Op. Cit., p.193

[3] ตรงตามความหมายของการตรวจสอบนัยสำคัญ $H_0 : \beta_j = 0$ VS $H_1 : \beta_j \neq 0$ คือปฏิเสธ H_0 (ยอมรับ H_1) ถ้า $F_c > F_{tab}$ = F - IN และยอมรับ H_0 ถ้า $Fc < F_{tab}$ = F - OUT

4.5.2.3 All possible regression

เนื่องจากในปัจจุบันนี้เทคโนโลยีด้านคอมพิวเตอร์ได้พัฒนาไปมาก คอมพิวเตอร์จึงเข้ามามีบทบาทในงานประมวลผลแทนเครื่องคิดเลข ทั้งมีโปรแกรมทางสถิติมากมายให้เลือกใช้ all possible regression ที่แต่เดิมถูกเพ่งมองว่าดี แต่ไม่น่าใช้ เพราะเสียเวลาวิเคราะห์ข้อมูลซ้ำซากและกินเวลาเกินควร กลับได้รับความสนใจแทน FS BE และ SW เพราะ all possible regression มีสมการให้เราพิจารณาคัดเลือกมากมาย ซึ่งเราสามารถใช้ทั้งเทคนิคทางสถิติและการวินิจฉัยเชิงอัตนิยมผสมกันไปได้ ไม่เหมือน BE FS และ SW ที่ค่อนข้างบังคับให้เลือกเอาจากตัวแบบที่ลดลงมาเป็นชั้นสุดท้ายเพียงตัวแบบเดียวหรือไม่กี่ตัวแบบ

ลองพิจารณากรณีสมการ $Y = f(X_2, X_3, X_4)$ จะเห็นว่าเราจะต้องวิเคราะห์โดยวิธี all possible regression ถึง $2^3 - 1 = 7$ สมการคือ

$$Y = \beta_1 + \beta_2 X_2 + u$$
$$Y = \beta_1 + \beta_3 X_3 + u$$
$$Y = \beta_1 + \beta_4 X_4 + u$$
$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$
$$Y = \beta_1 + \beta_2 X_2 + \beta_4 X_4 + u$$
$$Y = \beta_1 + \beta_3 X_3 + \beta_4 X_4 + u$$
$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

หากมีตัวแปรอิสระ 4 ตัว เราต้องวิเคราะห์สมการถดถอยถึง $2^4 - 1 = 15$ สมการ หากมีตัวแปรอิสระ 5 ตัว เราต้องวิเคราะห์สมการถดถอยถึง $2^5 - 1 = 31$ สมการ หากมีตัวแปรอิสระ 10 ตัว เราต้องวิเคราะห์สมการถดถอยถึง $2^{10} - 1 = 1023$ สมการ เป็นต้น ซึ่งนับว่าเป็นภาระมาก แต่เราก็สามารถใช้คอมพิวเตอร์ช่วยทุ่นแรงได้โดยไม่น่าจะต้องวิตกถึงภาระงานมากนัก สิ่งที่น่าสนใจหรือน่าวิตกมากกว่าภาระงานยังมีอยู่ สิ่งนั้นก็คือ เมื่อเรามีสมการมากมายมหาศาลเช่นนี้ จะคัดเลือกอย่างไรจึงจะรวดเร็ว

คำตอบก็คือให้ค่อย ๆ กรองเอาสมการที่ไม่เหมาะสมทิ้ง วิธีนี้เสียเวลามาก วิธีที่ดีวิธีหนึ่งคือ ให้ใช้ C_p – plot (เสนอโดย Mallows, C.L., 1973) หรือ C_p criterion วิธีนี้ให้คำนวณหาค่า C_p ของแต่ละสมการดังนี้

$$C_p = \frac{\text{residual sum square ของ subset model}}{s^2} - (n - 2p)$$

โดยที่ p คือจำนวนพารามิเตอร์ β ใน subset model n คือจำนวนค่าสังเกต s^2 คือ MSE ของสมการเต็มรูป (full model) ที่ประกอบด้วยตัวแปรอิสระครบ k – 1 ตัว หมายความว่า

เมื่อมีสมการ $2^{k-1} - 1$ สมการ เราก็จะได้ C_p ทั้งสิ้น $2^{k-1} - 1$ ค่า จากนั้นให้นำคู่ลำดับ (p, C_p) ไปพล็อตลงในระนาบที่แกนนอน คือค่า p ซึ่งมีค่าได้ตั้งแต่ 1 ถึง $k - 1$ ตามจำนวนตัวในสมการ เช่น ถ้ามีตัวแปร 3 ตัว ค่าของ p ใน subset model จะมีได้ 1, 2 และ 3 ตัว ค่า p จึงเท่ากับ 1, 2, 3 ส่วนแกนตั้งคือค่าของ C_p ตามตัวอย่าง subset model ของกรณีสมการ $Y = f(X_2, X_3, X_4)$ ข้างต้น เราจะมีคู่ลำดับดังนี้ คือ $(1, C_1), (1, C_1), (1, C_1), (2, C_2), (2, C_2), (2, C_2), (2, C_2), (3, C_3)$ ผลการพล็อตจะเห็นว่าจุดโคออร์ดิเนตกระจัดกระจาย โคออร์ดิเนตใดตกบนเส้นตรง 45° ให้ถือว่าเป็นโคออร์ดิเนตของสมการที่เป็น best subset model หากพบว่ามีมากกว่า 1 model ให้คัดเลือกโดยใช้เกณฑ์ร่วมอื่น ๆ เช่น R^2, s^2 รวมถึงความสอดคล้องหรือขัดแย้งกับข้อตกลงสมการถดถอย

หมายเหตุ เราสามารถคำนวณหาค่า C_p ได้อีกวิธีหนึ่งดังนี้ คือ

$$C_p = \frac{(n - k)(1 - R_1^2)}{1 - R^2} - (n - 2p)$$

โดยที่ R_1^2 คือค่าสัมประสิทธิ์การตัดสินใจจาก subset model และ R^2 คือค่าสัมประสิทธิ์การตัดสินใจจาก full model ความจริงแล้ววิธีนี้ก็คือวิธีเดียวกันกับวิธีที่ผ่านมา เพียงแต่เสนอสูตรไว้ในรูปที่ต่างกัน

ตัวอย่าง 4.4[1] จากการศึกษาของสตีนเนอร์และสตาร์คเรื่อง "ผลกระทบของส่วนผสมของซีเมนต์พอร์ตแลนด์ที่มีต่อการเปลี่ยนแปลงความร้อนขณะปูนแข็งตัว" โดยวัดปริมาณของส่วนผสมเป็นร้อยละต่อซีเมนต์ 1 หน่วย ดังนี้

X_1 = ร้อยละของไตรแคลเซี่ยมอลูมิเนต ($3\,CaO \cdot Al_2O_3$)

X_2 = ร้อยละของไตรแคลเซี่ยมซิลิเกต ($3\,CaO \cdot SiO_2$)

X_3 = ร้อยละของเตตระแคลเซี่ยมอลูมิโนเฟอร์ไรท์ ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$)

X_4 = ร้อยละของไดแคลเซี่ยมซิลิเกต ($2\,CaO \cdot SiO_2$)

Y = X_5 = ปริมาณความร้อนที่เปลี่ยนแปลงไปขณะปูนแข็งตัว วัดเป็นแคลอรี่ต่อปูน 1 กรัม

ผลการบันทึกข้อมูลปรากฏดังนี้

X_1	X_2	X_3	X_4	X_5 (หรือ Y)
7.0	26.0	6.0	60.0	78.5
1.0	29.0	15.0	52.0	74.3
11.0	56.0	8.0	20.0	104.3
11.0	31.0	8.0	47.0	87.6
7.0	52.0	6.0	33.0	95.9
11.0	55.0	9.0	22.0	109.2
3.0	71.0	17.0	6.0	102.7
1.0	31.0	22.0	44.0	72.5
2.0	54.0	18.0	22.0	93.1
21.0	47.0	4.0	26.0	115.9
1.0	40.0	23.0	34.0	83.8
11.0	66.0	9.0	12.0	113.3
10.0	68.0	8.0	12.0	109.4

จงวิเคราะห์สมการถดถอย $Y = f(X_1, X_2, X_3, X_4)$

[1] Drapper, N. and Smith, H., Op. Cit., p.365, 397-402

วิธีทำ การวิเคราะห์ข้อมูลโดยนัยของ Stepwise Regression ปรากฏดังต่อไปนี้ อย่างไรก็ตามก่อนที่นักศึกษาจะพบ Print Out ขออธิบายความหมายของข้อความบางส่วนที่ปรากฏใน Print Out ดังกล่าวเสียก่อนดังนี้

1. Original and / or Transformed Data คือข้อมูลที่ส่งเข้าสู่โปรแกรม Stepwise ข้อมูลนี้อาจเป็นข้อมูลลักษณะเดิมหรือข้อมูลที่แปลงรูปโดยเทคนิคการแปลงรูปลักษณะใดลักษณะหนึ่งตามเหตุผลและความจำเป็น ทั้งนี้ขึ้นอยู่กับผู้วิจัยจะเห็นสมควร เช่น

ก. $Y^{1/2}$, $X_j^{1/2}$ สำหรับบางค่าของ j ใช้ในสถานการณ์ที่ $\sigma_u^2 \propto E(Y_i)$ หรือ $\sigma_u^2 \propto X_{ji}$ โดยปกติแล้ว การแปลงรูปแบบนี้นิยมใช้ในกรณีที่ข้อมูลเดิมมีการแจกแจงแบบพัวซอง

ข. $Y^{1/2} + (Y+1)^{1/2}$, $X_j^{1/2} + (X_j+1)^{1/2}$ สำหรับบางค่าของ j ใช้ในสถานการณ์เดียวกับข้อ ก แต่ปรับปรุงให้เหมาะสมขึ้นเพื่อรับกับข้อมูลของตัวแปร X_j กับ Y ที่อาจมีบางส่วนมีค่าเท่ากับ 0 หรือใกล้ 0 (Freeman – Tukey Transformation)

ค. $\log(Y)$, $\log(X_j)$ สำหรับบางค่าของ j ใช้ได้สถานการณ์ที่ $\sigma_u^2 \propto [E(Y_i)]^2$ หรือ $\sigma_u^2 \propto X_{ji}^2$ และในกรณีที่ตัวแปรมีค่าอยู่ในช่วงค่อนข้างกว้าง หรือในกรณีต้องการให้ตัวแปรต่าง ๆ มีค่าอยู่ในระบบเดียวกัน เช่นแปลงข้อมูลของตัวแปรบางตัวที่วัดในรูปปริมาตรให้เป็นข้อมูลเชิงเส้นเช่นเดียวกับตัวแปรอื่น ๆ รวมตลอดถึงการใช้เพื่อแปลงรูปตัวแปรอันมีผลให้ตัวแปรสุ่ม u มีการแจกแจงคืนสู่การแจกแจงปกติ[1] แต่ทั้งนี้ตัวแปรทุกตัวต้องมีค่าเป็นบวกเสมอ

ง. $\log(Y+1)$, $\log(X_j+1)$ สำหรับบางค่าของ j ใช้ได้ในสถานการณ์เดียวกับข้อ ค. แต่ปรับปรุงให้เหมาะสมกับสถานการณ์ที่ตัวแปร X_j และ Y มีค่าบางค่าเป็น 0

จ. $\frac{1}{Y}$, $\frac{1}{X_j}$ สำหรับบางค่าของ j ใช้ได้ในสถานการณ์ $\sigma_u^2 \propto [E(Y_i)]^4$ หรือ $\sigma_u^2 \propto X_{ji}^4$ และในกรณีที่ค่าของตัวแปรเกาะกลุ่มกันใกล้ 0 การแปลงรูปลักษณะนี้นิยมใช้ในกรณีของตัวแปรที่เกี่ยวข้องกับเวลารอคอย อัตราการอยู่รอด ผลการรักษาโรค และอื่น ๆที่เกี่ยวข้องอยู่กับตัวแปรเวลา

ฉ. $\frac{1}{Y+1}$, $\frac{1}{X_j+1}$ สำหรับบางค่าของ j ใช้ได้เช่นเดียวกับข้อ ง. แต่ปรับปรุงให้สอดคล้องกับสถานการณ์ของตัวแปรมีค่าบางค่าเท่ากับ 0

ตัวแปรในข้อ ง และ ฉ ต้องเป็นจำนวนที่ไม่ติดลบ (Nonnegative)

ช. $\text{Sin}^{-1}(\sqrt{Y})$, $\text{Sin}^{-1}(\sqrt{X_j})$ สำหรับบางค่าของ j กรณีนี้ใช้ในสถานการณ์ที่ $\sigma_u^2 \propto E(Y_i)(1-E(Y_i))$ หรือ $\sigma_u^2 \propto [X_{ji}(1-X_{ji})]$ และในกรณีที่ตัวแปรมีค่าในรูปของสัดส่วนคือ $0 \leqslant Y_i \leqslant 1$, $0 \leqslant X_{ji} \leqslant 1$

---

[1] Maddala, G.S., Op. Cit., p.314-316

2. Means of Transformed Variables คือค่าเฉลี่ยของตัวแปร

3. Std. deviations of Transform Variables คือส่วนเบี่ยงเบนมาตรฐานของตัวแปร

4. Correlation matrix คือเมตริกซ์ของค่าสหสัมพันธ์เชิงเส้น ทั้งระหว่างตัวแปรอิสระด้วยกัน และระหว่างตัวแปรตามกับตัวแปรอิสระ สมมุติว่ามีตัวแปรอิสระ 4 ตัว ลักษณะการเสนอแมตริกซ์จะปรากฏในรูปแบบดังนี้

	X_1	X_2	X_3	X_4	$X_5(=Y)$
X_1	1	r_{12}	r_{13}	r_{14}	r_{15}
X_2	r_{21}	1	r_{23}	r_{24}	r_{25}
X_3	r_{31}	r_{32}	1	r_{34}	r_{35}
X_4	r_{41}	r_{42}	r_{43}	1	r_{45}
X_5	r_{51}	r_{52}	r_{53}	r_{54}	1

ค่า r_{ij} เมื่อ $i = 1, 2, 3, 4$ และ $j = 1, 2, 3, 4$ คือค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ ในขณะที่ r_{i5} ; $i = 1, 2, \ldots, 5$ และ r_{5j} ; $j = 1, 2, \ldots, 5$ คือค่าสหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม

5. Control Information ประกอบด้วย

ก. No. of Observations คือจำนวนค่าสังเกตหรือขนาดตัวอย่าง

ข. Response Variable is no. ข้อความนี้ใช้เพื่อชี้หรือสอบย้ำให้ทราบว่าตัวแปรตัวใดหรือนัยหนึ่ง สดมภ์ที่เท่าไร ของ Input Matrix คือตัวแปร Y โดยปกติโปรแกรม Stepwise จะใช้สดมภ์สุดท้ายเป็นสดมภ์ของค่าตัวแปรตามเสมอ

ค. Risk level for B confidence interval คือ α - error.

ง. F level for entering a variable คือ F-IN ที่กล่าวมาแล้วในตอนต้น

จ. F level for deleting a variable คือ F-OUT ที่กล่าวมาแล้ว โดยปกติทั้ง F-IN และ F-OUT ไม่จำเป็นต้องมีค่าเท่ากัน แต่อาจกำหนดให้เท่ากันก็ได้ทั้งนี้ขึ้นอยู่กับระดับนัยสำคัญ α ที่นักวิจัยจะกำหนดขึ้นตามระดับที่ตนเองเห็นว่าเหมาะสมหรือพอจะยอมรับได้

ช. Variable entering เป็นข้อความที่ใช้ระบุว่า ณ ขั้นตอนนั้น ๆ (Step. No.) ตัวแปรอิสระตัวใดได้รับการคัดเลือกเข้าสู่สมการ เช่น ในขั้นที่ 1 Variable Entering คือ 4 แสดงว่า ในขั้นแรกนี้ตัวแปรอิสระที่เข้าสู่สมการก่อนตัวอื่น คือ X_4

ซ. Sequential F-test คือค่า F ที่ใช้ทดสอบดูว่าตัวแปรอิสระที่เข้าสู่สมการในขั้นตอนนั้น ๆ (ถือว่าเป็นตัวแปรอิสระที่เข้าสู่สมการเป็นตัวสุดท้ายเสมอ แม้ว่าจะเป็น Step ที่ 1 ก็ตาม)

มีนัยสำคัญหรือไม่ หรือนัยหนึ่งก็คือ เป็นค่า F ที่ใช้ทดสอบดูว่าตัวแปรล่าสุดที่เข้าสู่สมการนั้นสามารถช่วยลดปริมาณของ Unexplained Variation ($\sum e_i^2$) ลงได้อย่างมีนัยสำคัญหรือไม่ สมมุติฐานสำหรับกรณีนี้ก็คือ $H_0 : \beta_\ell = 0$ VS $H_1 : \beta_\ell \neq 0$ เมื่อ ℓ คือดัชนี หรือ Subscript ของ X_ℓ หรือ X ที่เข้าสู่สมการเป็นตัวสุดท้ายในขั้นตอนนั้น

ฌ. Percent variation Explained คือ R - SQ หรือ R^2 หรือ $(\text{Multiple Correlation})^2$

ญ. Mean of the Response คือค่าเฉลี่ยของตัวแปร Y

ฏ. Std. error as a % of mean response คือ $\frac{\hat{\sigma}}{\bar{Y}} \times 100\ \%$ โดยที่ $\hat{\sigma}$ คือ Standard error of Residual ของสมการในแต่ละขั้น

$\frac{\hat{\sigma}}{\bar{Y}}$ คือCoefficient of Variation (C.V.) ใช้เป็นดัชนีสำหรับเปรียบเทียบ Relative Dispersionในแต่ละ step การเปรียบเทียบค่า C.V. ให้พิจารณาที่ขนาดหรือปริมาณของ C.V. ในแต่ละ Step ถ้าค่า C.V. ใน Step ใดสูงกว่า แสดงว่าความผันแปร (Unexplained Variation) จะสูงกว่าหรือสมการใน Step นั้นยังใช้พยากรณ์ไม่ได้ดีเท่ากับสมการใน Step ที่ให้ค่า C.V. ต่ำกว่า[1]

---

[1] ดัชนีที่สมควรปรากฏอยู่ใน Control Information แต่กลับไม่ปรากฏในโปรแกรมคือ Coefficient of Skewness และ Coefficient of Kurtosis (Flatness) ซึ่งสามารถใช้วัดดูว่า u มีการแจกแจงสมมาตรและเป็นแบบปกติหรือไม่ นิยามของสัมประสิทธิ์ทั้งสองปรากฏดังนี้

$$\text{Coefficient of Skewness } \gamma_1 = \frac{\mu_x^3}{\sigma_x^3} \text{ และ } \hat{\gamma}_1 = \frac{(1/n)\sum(x_i - \bar{x})^3}{s^3}$$

$$\text{Coefficient of kurtosis } \gamma_2 = \frac{\mu_x^4}{\sigma_x^4} \text{ และ } \hat{\gamma}_2 = \frac{(1/n)\sum(x_i - \bar{x})^4}{s^4}$$

การตัดสินใจให้กระทำดังนี้

ก. ถ้า $\hat{\gamma}_1 > 0$ แสดงว่าโค้งเบ้ขวา ถ้า $\hat{\gamma}_1 < 0$ แสดงว่าโค้งเบ้ซ้าย ถ้า $\gamma_1 \cong 0$ แสดงว่าโค้งสมมาตร

ข. ในกรณีของตัวแปรสุ่มที่มีการแจกแจงปกติ ค่า $\hat{\gamma}_2$ จะเท่ากับ 3.0 โดยปกติจะถือว่า $\hat{\gamma}_2 = 3.0$ คือค่ามาตรฐาน ดังนั้นถ้าตัวแปรสุ่มใดให้ค่า $\hat{\gamma}_2 \cong 3.0$ แสดงว่าตัวแปรสุ่มนั้นมีการแจกแจงแบบปกติ ถ้า $\hat{\gamma}_2 > 3.0$ แสดงว่าตัวแปรสุ่มนั้นมีโค้งการแจกแจงความถี่แบนหรือต่ำกว่ามาตรฐาน ถ้า $\hat{\gamma}_2 < 3.0$ แสดงว่าตัวแปรสุ่มนั้นมีการแจกแจงความถี่โด่งกว่ามาตรฐาน โดยปกติเรานิยมที่จะใช้ Coefficient of Excess คือ $\hat{\gamma}_2 - 3$ เป็นดัชนีเปรียบเทียบว่าโค้งที่เราสนใจแบนหรือโด่งกว่าโค้งปกติมากกว่าที่จะใช้ Coefficient of Kurtosis กล่าวคือ ถ้า $(\hat{\gamma}_2 - 3) > 0$ แสดงโค้งที่สนใจแบนต่ำกว่าโค้งปกติ และ ถ้า $(\hat{\gamma}_2 - 3) < 0$ แสดงว่าโค้งที่สนใจโด่งกว่าโค้งปกติ

สำหรับกรณีของ Stepwise Regression นักวิจัยสามารถใช้ข้อมูลของ e หรือ Y มาคำนวณหาสัมประสิทธิ์ $\hat{\gamma}_1$, $\hat{\gamma}_2$ และ $(\hat{\gamma}_2 - 3)$ ได้

ฎ. Degree of Fredom หมายถึง df ของ Residual Source.

ฐ. Deteminant Value หมายถึง ดีเทอร์มิแนนต์ของ Correlation Matrix ตามแบบใน ข้อ 4. สำหรับในแต่ละ Step นั้นสมาชิกของ Correlation Matrix จะประกอบไปด้วยเฉพาะค่าสหสัมพันธ์ระหว่างตัวแปรอิสระเฉพาะใน Step นั้น ๆ เท่านั้น

นั่นคือ Deleminant Value = $|R_x|$ โดยที่ R_x คือ Correlation Matrix ที่ประกอบด้วยสมาชิกที่แสดงสหสัมพันธ์ระหว่างตัวแปรอิสระในแต่ละ Step.

$|R_x|$ มีประโยชน์ในเชิงวิเคราะห์ ดังนี้

(1) ใช้ในการคำนวณหา R_x^{-1} เพื่อนำไปสู่การประมาณค่าเวคเตอร์ β

(2) ใช้เป็นเกณฑ์ในการพิจารณาว่ามีปัญหา Multicolinearity เกิดขึ้นหรือไม่

สมมุติว่าตัวแปรอิสระ X_1 และ X_2 มีความสัมพันธ์เชิงเส้นต่อกัน กล่าวคือ $X_1 = cX_2$ ซึ่งในกรณีนี้จะพบว่า

$$r_{12} = \frac{\sum_i^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_i^n (X_{1i} - \bar{X}_1)^2 \sum_i^n (X_{2i} - X_2)^2}} = \frac{c\sum_i^n (X_{2i} - \bar{X}^2)^2}{\sqrt{c_i^2 (\sum_i^n (X_{2i} - \bar{X}_1)^2)^2}} = 1$$

ดังนั้น $R_x = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ และพบว่า $|R_x| = 0$

ดังนี้ย่อมแสดงว่า ถ้า $|R_x| = 0$ ตัวอิสระย่อมมีความสัมพันธ์ (Colinear) ต่อกันอย่างสมบูรณ์ และถ้า $|R_x| \to 0$ ก็แสดงว่าตัวแปรอิสระมีความสัมพันธ์ต่อกันในอัตราที่ค่อนข้างสูง โดยนัยกลับกัน ถ้า $r_{12} = 0$ หรือ X_1 และ X_2 เป็นอิสระเชิงเส้นต่อกัน เราจะพบว่า

$$|R_x| = \begin{vmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$$

ซึ่งแสดงว่า ถ้า $|R_x| = 1$ หรือใกล้ 1 ย่อมเป็นดัชนี ชี้ให้เห็นว่าตัวแปรอิสระทั้งหลายเป็นอิสระหรือค่อนข้างอิสระต่อกันและกัน

ดังนั้น Determinant Value หรือ $|R_x|$ จึงใช้เป็นเครื่องมือที่ช่วยชี้ให้เห็นได้อย่างหยาบ ๆ ว่าตัวแปรอิสระที่เข้าสู่สมการแล้วใน Step ต่าง ๆ มีความเป็นอิสระต่อกันมากน้อยเพียงใด โดยที่ $0 \leq |R_x| \leq 1$ สำหรับเครื่องมือที่ใช้ทดสอบว่าตัวแปรอิสระมีความเป็นอิสระต่อกัน (Orthogonal

หรือ Independent) อย่างมีนัยสำคัญหรือไม่นั้นได้มีผู้เสนอไว้แล้วคือฟาราห์และกลอเบอร์ (D.E. Farrar and R.R. Glauber , 1967)[1] ฟาร์ราห์และกลอเบอร์เสนอตัวทดสอบสำหรับสมมุติฐาน

H_0 : X's เป็นอิสระต่อกัน VS H_1 : X's ไม่เป็นอิสระต่อกันดังนี้คือ

ปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α ถ้า $*\chi^2 > \chi^2_\nu$ โดยที่

$*\chi^2 = -[n-1-\frac{1}{6}(2k+5)\log_e|R_x|]$ เมื่อ k = จำนวนตัวแปรอิสระ ในกรณีของ Stepwise Regression k หมายถึงจำนวนตัวแปรอิสระที่เข้าสู่สมการแล้วในแต่ละ Step

χ^2_ν คือค่าวิกฤตจากตาราง χ^2 เมื่อ $\nu = \frac{1}{2}k(k-1)$

6. B Coefficient and Confidence Limits

ก. Var No. แสดงเลขที่หรือเลขประจำตัวของตัวแปรที่เข้าสู่สมการ

ข. Mean หมายถึงค่าเฉลี่ยของข้อมูลของตัวแปรอิสระ

ค. Decoded B Coefficient หมายถึงสัมประสิทธิ์ความถดถอย โดยปกติจะเป็นค่าประมาณของ β_j ; j = 1, 2, ... , k โดยที่ β_j คือ พารามิเตอร์ของสมการถดถอย แต่ถ้าข้อมูลได้รับการแปลงรูปด้วยเทคนิคใดเทคนิคหนึ่ง (ซึ่งต้องระบุให้ชัดเจนด้วย) แล้วโปรแกรมจะแปลงรูป $\hat{\beta}_j$ ให้สอดคล้องกับข้อมูลเดิม ตัวอย่างเช่น สมการเดิมคือ $Y = \beta_1 X_2^{\beta_2} u$ ถ้าเราแปลงรูปสมการนี้ให้เป็นสมการเชิงเส้น โดยอาศัย Log Transformation คือ

$\log Y = \log \beta_1 + \beta_2 \log X_2 + \log u$ ซึ่งในกรณีเช่นนี้ค่าประมาณของพารามิเตอร์ที่ได้คือ $\widehat{\log \beta_1}$ และ $\hat{\beta}_2$ ซึ่งโปรแกรมจะแปลงรูปจาก $\widehat{\log \beta_1}$ เป็น $\hat{\beta}_1$ ด้วยการใช้ Antilog

อย่างไรก็ตามเรื่องการแปลงรูปย้อนกลับสู่รูปเดิมไม่ใช่สิ่งจำเป็นเพราะไม่มีผลต่อคุณภาพของงานแต่ประการใด เพราะสมการเดิมหรือสมการที่แปลงรูปแล้วต่างก็ใช้ในการพยากรณ์ได้ผลตรงกัน ด้วยเหตุนี้เพื่อป้องกันความสับสน ผู้วิจัยพึงแปลงรูปข้อมูลสู่รูปที่ต้องการให้เรียบร้อยเสียก่อนแล้วจึงบันทึกลงตัวกลางข้อมูล ผลลัพธ์คือ $\hat{\beta}_j$ จะเป็นค่าประมาณของ β_j จากสมการที่ผู้วิจัยตัดสินใจเลือกเอาไว้ด้วยเหตุนี้ผู้เขียนจึงกล่าวว่า Decoded B Coefficient นั้นโดยปกติก็หมายถึงค่าประมาณของ β_j ตามปกตินั่นเอง

J. Limits Upper/Lower หมายถึงช่วงเชื่อมั่น 95% (Random Interval) ที่คาดว่าค่าจริงของ β_j จะปรากฏอยู่ การคำนวณหาช่วงเชื่อมั่นนี้ใช้สูตรดังนี้คือ

---

[1] Koutsoyiannis , A. , Op. cit., p.236

$$P_r\{\hat{\beta}_j - \hat{\sigma}_{\hat{\beta}_j} t_{n-k}, \frac{\alpha}{2} \leqslant \beta_j \leqslant \hat{\beta}_j + \hat{\sigma}_{\hat{\beta}_j} t_{n-k}, 1-\frac{\alpha}{2}\} = 1-\alpha$$

โดยที่ k = จำนวนพารามิเตอร์ β ในสมการถดถอยเฉพาะ step นั้น ๆ

จ. Standard error หมายถึง Standard error ของ $\hat{\beta}_j$

ฉ. Partial F - test คือค่า F สำหรับทดสอบนัยสำคัญของตัวแปรอิสระแต่ละตัว (ความจริงก็คือทดสอบนัยสำคัญของพารามิเตอร์ β ที่คู่อยู่กับตัวแปรอิสระนั้น ๆ) โดยถือว่าตัวแปรอิสระแต่ละตัวเข้าสู่สมการเป็นตัวสุดท้ายหมายความว่า การคำนวณหา Partial F นักวิจัยจะต้องสลับที่ตัวแปรอิสระเพื่อให้ตัวแปรอิสระทุกตัวได้รับการจัดลำดับให้เข้าสู่สมการเป็นตัวล่าสุดจึงจะคำนวณหาค่า Partial F ได้

เช่นใน Step ที่ 3 มีตัวแปรอิสระเข้าสู่สมการ 3 ตัว คือ X_2, X_4, X_1 นักวิจัยจะต้องคำนวณหาค่า Partial F สำหรับ β_2, β_4, และ β_1 ซึ่งสามารถกระทำได้ดังนี้

(1) จาก $Y = f(X_2, X_4, X_1)$ คำนวณหา $SS(\beta_0, \beta_2, \beta_4, \beta_1,)$ และ $\hat{\sigma}^2 = s^2$

(2) จาก $Y = f(X_2, X_4)$ คำนวณหา $SS(\beta_0, \beta_2, \beta_4)$

(3) จาก $Y = f(X_2, X_1)$ คำนวณหา $SS(\beta_0, \beta_2, \beta_1)$

(4) จาก $Y = f(X_4, X_1)$ คำนวณหา $SS(\beta_0, \beta_4, \beta_1)$

(5) จาก (1) และ (2) Partial F สำหรับ β_1 คือ $\dfrac{SS(\beta_0, \beta_2, \beta_4, \beta_1) - SS(\beta_0, \beta_2, \beta_4)}{s^2}$

(6) จาก (1) และ (3) Partial F สำหรับ β_4 คือ $\dfrac{SS(\beta_0, \beta_2, \beta_4, \beta_1) - SS(\beta_0, \beta_2, \beta_1)}{s^2}$

(7) จาก (1) และ (4) Partial F สำหรับ β_2 คือ $\dfrac{SS(\beta_0, \beta_2, \beta_4, \beta_1) - SS(\beta_0, \beta_4, \beta_1)}{s^2}$

ช. Constant terms in prediction equation หมายถึง $\hat{\beta}_0$ สำหรับรูปแบบสมการที่พัฒนาในหนังสือเล่มนี้ผู้เขียนใช้ β_1 หมายถึง เทอมคงที่

ซ. Square of partial Correlation Coefficient of Variables not in regression หมายถึง $R^2_{1j \cdot 23 \ldots j-1, j+1, \ldots, k}$ เมื่อ Subscipt j แสดงลำดับหรือเลขประจำตัวของตัวแปรอิสระที่ยังไม่เข้าสู่สมการ

เช่น สมการ $Y = f(X_2, X_3, X_4, X_5, X_6,)$ และ X_2 กับ X_3 ได้รับการพิจารณานำเข้าสู่สมการแล้วใน Step ที่ 2 ดังนั้น Square of partial Correlation จะประกอบ $R^2_{1j \cdot 23}$ โดยที่ $j = 4, 5, 6$ ...

ฌ. Normal Diviate หมายถึง $\frac{e_i}{s} = \frac{Y_i - \hat{Y}_i}{s}$ โดยที่ s หรือ $\hat{\sigma}_u$ คือ Standard error of residual กล่าวคือ

s^2 หรือ $\hat{\sigma}_u^2 = \frac{1}{n-k}\sum_i^n e_i^2$ และ s หรือ $\hat{\sigma}_u = \sqrt{\sum_i^n e_i^2 / n-k}$ เมื่อ k คือจำนวนพารามิเตอร์ของสมการถดถอย

4.6 การตรวจสอบความเหมาะสมของสมการและเทคนิคที่น่าสนใจ

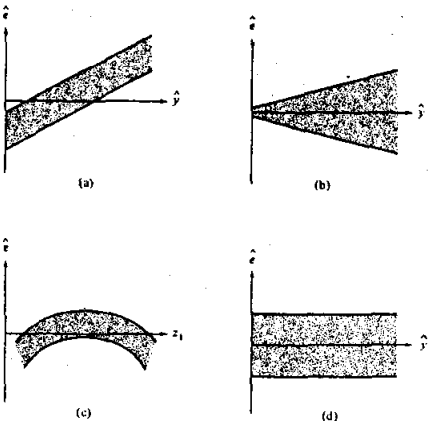
4.6.1 การตรวจสอบความเหมาะสมของสมการ[1/]

สมมุติว่าสมการ $Y = X\hat{\beta}$ ถูกต้องแล้ว ก่อนที่เราจะนำสมการนี้ไปใช้งานจริงเราควรจะได้ตรวจสอบความเหมาะสม (fitness) เสียก่อน วิธีปฏิบัติที่เรานิยมใช้คือการนำเอา e_i ไปพล็อตด้วยวิธีต่างๆ เพื่อตรวจดูว่า u_i สอดคล้องกับข้อตกลงหรือไม่ ควรปรับปรุงสมการอีกต่อไปหรือไม่

<u>วิธีที่ 1</u> การนำ e_i ไปพล็อตร่วมกับ $\hat{Y}_i$; i = 1,2,..., n ผลการพล็อตจะฟ้องถึงความไม่เหมาะสมเป็น 2 แนวดังนี้

(1) ถ้ากลุ่มคู่ลำดับ ($\hat{Y}_i$, e_i) เกาะกลุ่มเป็นแนวดังภาพ (a) แสดงว่า e_i ผันแปรไปตาม $\hat{Y}_i$ ซึ่งชี้ว่าเราจำเป็นต้องผนวกตัวแปรอิสระ X_j อื่นๆ เพิ่มลงในสมการ Y = f(X' s) เดิมอีกจำนวนหนึ่งหรืออาจต้องแปลงรูปตัวแปรหรือสมการ หรือทั้งสองประการ

(2) ถ้ากลุ่มคู่ลำดับ ($\hat{Y}_i$, e_i) เกาะกลุ่มเป็นแนวดังภาพ (b) แสดงว่า σ^2 (ซึ่งก็คือ $V(Y_i)$) มีค่าเพิ่มขึ้นตามค่า $\hat{Y}_i$ และลดลงตามค่า $\hat{Y}_i$ แสดงว่า σ^2 ไม่มีลักษณะเป็นค่าคงที่ตามข้อตกลงว่า



---

[1] เรื่องนี้จะกล่าวถึงเพียงเล็กน้อยพอให้เห็นแนวทางการดีความซึ่งโปรแกรมสำเร็จรูปส่วนใหญ่เช่นTSPหรือSPSS นิยมปฏิบัติ ความจริงแล้วเรื่องเหล่านี้ได้กล่าวไว้โดยละเอียดในบทที่ 5-12

$E(u_i^2) = \sigma^2$ (ปัญหา heteroscedastic disturbance) ต้องประมาณค่าเวคเตอร์ตามวิธี GLS (อ่านบทที่ 6) ถ้าคู่ลำดับเกาะกลุ่มกันดังภาพ (d) แสดงว่า σ^2 มีค่าคงที่

วิธีที่ 2 การพล็อต e_i ร่วมกับตัวแปรอิสระหรือฟังก์ชันของตัวแปรอิสระ ถ้าคู่ลำดับ $(e_i, f(Xj))$ เกาะกันเป็นแนวดังภาพ (d) แสดงว่าเราต้องผนวก f(Xj) ลงในสมการ

วิธีที่ 3 Q-Q plot คือการนำเอา observed quantile ซึ่งก็คือ $e_{(i)}$ ($e_{(i)}$ เป็น order statistics) ไปพล็อตร่วมกับ expected quantile คือ $q_{(i)}$ ซึ่งโดยปกติจะถือว่าเป็น quantile ของ standard normal density ถ้ากลุ่มของคู่ลำดับ $(q_{(i)}, e_{(i)})$ เกาะกันเป็นแนวเส้นตรง แสดงว่าข้อตกลงว่าตัวแปรสุ่ม u แจกแจงแบบปกติเป็นจริง (อ่านบทที่ 5)

quantile $q_{(i)}$ ก็คือค่าคงที่ (ordinate) บนแกนนอนใต้โค้ง f ที่มีผลให้พื้นที่ใต้โค้งที่สะสมจากปลายซ้ายมาจนถึง q มีค่าเท่ากับส่วนร้อยที่เรากำหนด (คือ $p_{(i)}$)

ถ้าให้ส่วนร้อย (หรือความน่าจะเป็น) คือ $p_{(i)}$ และ f คือ N(0,1) เรานิยาม quantile $q_{(i)}$ ได้ดังนี้

$$\Pr\{Z \leq q(i)\} = \int_{-\alpha}^{q(i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

เช่นถ้ากำหนดให้ $p_{(i)} = .05$ เราย่อมหา $q_{(i)}$ ได้โดยง่ายคือ

$$\Pr\{Z \leq -1.645\} = .05 \text{ แสดงว่า } q_{(i)} = -1.645$$

ขอให้สังเกตว่าการใช้ Q-Q plot นั้นเรามีจุดประสงค์เพื่อตรวจสอบข้อตกลงเรื่อง normality ดังนั้นเราจึงกำหนดให้ f เป็น N(0,1) ค่า quantile จึงหมายถึง ordinate ใต้โค้ง N(0,1) นั่นเอง

ในทางปฏิบัติเราจะดำเนินการเพื่อทำ Q-Q plot เป็นขั้นๆ ดังนี้

ขั้นที่ 1 วิเคราะห์สมการถดถอย $Y = X\beta + U$ แล้วนำเอา residual e มาใช้งาน โดยที่ $e = [e_1, e_2, ..., e_n]$

ขั้นที่ 2 เรียงลำดับค่าของ $e_1, e_2, ..., e_n$ ตามลำดับจากน้อยไปหามากได้ $e_{(1)} < e_{(2)} < ... < e_{(i)} < ... < e_{(n)}$

ขั้นที่ 3 จาก ordered residual $e_{(1)} < e_{(2)} < \ldots < e_{(i)} < \ldots < e_{(n)}$ ให้คำนวณหาความถี่สะสมเฉพาะตัวของ $e_{(i)}$ ซึ่งเราจะใช้เป็นค่าประมาณของความน่าจะเป็นต่อไปได้ดังนี้คือ

$$\text{ความน่าจะเป็น} = \frac{i}{n} = p_{(i)}$$

ด้วยเหตุนี้ความน่าจะเป็น $p_{(i)}$ ของแต่ละ $e_{(i)}$ จึงปรากฏดังนี้คือ

$$\frac{1}{n} < \frac{2}{n} < \ldots < \frac{1}{n} < \ldots < \frac{n}{n} \text{ ซึ่งก็คือ } p_{(1)} < p_{(2)} < \ldots < p_{(i)} < \ldots < p_{(n)}$$

จากนี้ให้ปรับค่า $p_{(i)}$ เล็กน้อยเพื่อให้เกิดความต่อเนื่อง เรียกว่า continuity correction ได้ p(i) ของแต่ละ e(i) ดังนี้ [1/]

$$(1-\tfrac{1}{2})/n\ ,\ (2-\tfrac{1}{2})/n < \ldots < (i-\tfrac{1}{2})/n < \ldots > (n-\tfrac{1}{2})/n$$

หรืออาจใช้เป็น

$$(1-\tfrac{3}{8})/(n+\tfrac{1}{4}) < (2-\tfrac{3}{8})/(n+\tfrac{1}{4}) < \ldots < (i-\tfrac{3}{8})/(n+\tfrac{1}{4}) < \ldots < (n-\tfrac{3}{8})/(n+\tfrac{1}{4})$$

ขั้นที่ 4 จาก $p_{(i)}$ ในขั้นที่ 3 ให้คำนวณหา $q_{(i)}$ โดยอาศัยตาราง N(0,1) คือ $p_{(i)} = \Pr\{Z \leq q_{(i)}\}$; i = 1,2, ...,n เราจะได้ค่า $q_{(i)}$ ตามต้องการคือ $q_{(1)} < q_{(2)} < \ldots < q_{(i)} < \ldots < q_{(n)}$

ขั้นที่ 5 นำคู่ลำดับ $(q_{(i)}, e_{(i)})$ ไปพล็อตในระนาบที่มี q เป็นแกนนอนและมี e เป็นแกนตั้ง ถ้าคู่ลำดับเหล่านี้เกาะแนวกันเป็นเส้นตรงแสดงว่า u มีการแจกแจงแบบปกติ ถ้าไม่เป็นเส้นตรงแสดงว่าเกิดปัญหา nonnormality

วิธีที่ 4 การพล็อต e_i ร่วมกับลำดับกาล (t) ใช้สำหรับตรวจดูว่า มีปัญหา auto-correlation หรือไม่ ขณะเดียวกันเรายังสามารถใช้ตรวจดูได้ด้วยว่า σ^2 ผันแปรไปตามกาลเวลาหรือไม่ และควรผนวกตัวแปรเวลา t เข้ามาในสมการหรือไม่ ถ้าควรเราจะต้องใช้ t รูปใด t หรือว่า t^2 หรือว่า t^3 กล่าวคือถ้าคู่ลำดับ (t, e_t) ในที่นี้ใช้ e_t แทน e_i คล้ายภาพ (a) แสดงว่าเราควรเพิ่ม

---

[1] Jolnson, R.A. and D.W. Wichem, Applied Multivariate Statistical Analysis, (Prentice - Hall, Inc., NJ., 1982), p. 152.

t เป็นตัวแปรอิสระลงในสมการ ถ้าปรากฏคล้ายภาพ (b) แสดงว่า σ^2 เป็นฟังก์ชันของ t (คือ σ^2 ผันแปรตามกาล) ถ้าปรากฏคล้ายภาพ (c) แสดงว่าเราควรเพิ่มทั้ง t และ t^2 เป็นตัวแปรอิสระลงในสมการ ถ้าปรากฏคล้ายภาพ (d) แสดงว่ากาล (t) ไม่มีผลกระทบต่อข้อมูล

สำหรับการตรวจสอบ autocorrelation เราน่าจะใช้วิธีที่ดีกว่าการพล็อต (อ่านบทที่ 7)

วิธีที่ 5. outliers เป็นการนำเอา e_i ไปพล็อตตามลำดับค่าโดยที่แกนนอนคือค่าของ e_i แกนตั้งคือค่าของ i

เนื่องจาก $u_i \sim N\ (0,\ \sigma^2)$ ดังนั้น $u_i/\sigma - N(0,1)$; $i = 1,2,...,$ n เมื่อเราเกี่ยวข้อง กับ e_i ซึ่งเป็นค่าประมาณของ u_i เราจึงใช้ s^2 เป็นค่าประมาณของ σ^2 โดยที่

$$s^2 = \frac{1}{n-k} \sum_{i}^{n} (e_i - e)^2 = \frac{1}{n-k} \sum_{i}^{n} e^2 i$$

ด้วยเหตุนี้ e_i/s จึงเป็น Unit normal deviated residual

การทำ residual plot เพื่อหา ontlier ปฏิบัติได้ดังนี้

ขั้นที่ 1 แปลง e_i เป็น e_i/s ; i = 1,2,..., n

ขั้นที่ 2 คำนวณหา confident band ขั้นนี้จะพบว่า 95% confident band คือ [-1.96, 1.96] หรือ 99% confident band คือ [-2.58, 2.58]

ขั้นที่ 3 นำ $e_{i/S}$ ไปพล็อตลงในระนาบที่กำหนด confident band ค่า $e_{i/S}$ ที่ตกอยู่นอก confident band เรียกว่า outlier

ขั้นที่ 4 เพ่งเล็งค่าสังเกตที่สอดคล้องกับ outlier เป็นพิเศษโดยเราอาจเลือกปฏิบัติต่อค่าสังเกตเหล่านี้ได้ 2 แนวคือ ตัดค่าสังเกตนั้นทิ้งด้วยเหตุผลว่าเป็นค่าสังเกตที่ผิดปกติ หรือพินิจพิเคราะห์ค่าสังเกตนั้นอย่างถี่ถ้วนว่าเพราะเหตุใดจึงผิดเพี้ยนไปจากหมู่คณะ ซึ่งเราอาจพบความจริงอื่นๆ ที่น่าสนใจ เช่น สภาพแวดล้อมผิดแผกไปจากปกติ จึงทำให้ค่าสังเกตนั้นเป็น outlier ซึ่งหากมีเหตุผลดีพอเราจะไม่ตัดค่าสังเกตนี้ทิ้ง เพราะถือว่าให้คุณมากกว่าให้โทษ

วิธีที่ 6. การทดสอบ lack of fit เรื่องนี้จะกระทำเมื่อมี replication ของตัวแปรอิสระ กล่าวคือถ้าเรามีการทดลองซ้ำๆ ณ ค่า Xi เราย่อมได้ค่า Y หลายค่า (เท่ากับจำนวนซ้ำ) เช่น Xi = 1= จำนวนปุ๋ยที่ใส่ลงในแปลงทดลอง โดยเราจะใส่ปุ๋ยในแปลงทดลองขนาดมาตรฐานแปลงละ 1 กิโลกรัม รวม 100 แปลง (คือ 100 ซ้ำ) ผลคือได้ค่าผลผลิตผัก 100 ค่าซึ่งอาจเท่ากันหรือ

ไม่เท่ากันก็ได้ แต่โดยปกติจะไม่เท่ากันเพียงแต่ใกล้เคียงกัน ความผันแปรของ Y ณ X = 1 เรียกว่า pure error ณ X = 1 ถ้าเปลี่ยนค่าของ X ไปสู่ค่าอื่นและทำ replication ณ ค่านั้น เราก็ย่อมหาค่า pure error ณ ค่านั้นของ X เมื่อรวมความผันแปรทั้งปวงเหล่านี้เข้าด้วยกันก็จะได้ pure error sum square และได้ pure error mean square คือ s_e^2

$$s_e^2 = \frac{1}{[(n-1)} [\sum_{u}^{n_i} (y_{1u} - \bar{y}_1)^2 + \sum_{u}^{m} (y_{2u} - \bar{y}_2)^2 + ... + \sum_{u}^{nm} (y_{mu} - \bar{y}_m)^2]$$

โดยที่ n_i คือจำนวน repeication ณ จุดที่ $X_i = X_{io}$

เราถือว่า s^2e คือค่าประมาณของ σ^2

SS(lack of fit) = SS(residual) - SS(pure error)

$$df = (n-k) - \sum_{i=1}^{m} (n_i - 1) = \nu$$

MS(lack of fit) = SS(lack of fit)/ν

การทดสอบ lack of fit คือ Ho : no lack of fit VS H_1 : lack of fit เราตัดสินใจดังนี้

ปฏิเสธ Ho เมื่อ $MS(\text{lack of fit})/s_e^2 > F_{\nu, \Sigma(n_i-1), 1-\alpha}$

แปลว่า ถ้า Ho มีนัยสำคัญ (คือปฏิเสธ Ho) เราก็สรุปได้ว่าสมการถดถอยที่ได้ในขณะนั้นยังไม่เหมาะสม กรณีนี้เราจะต้องเปลี่ยนรูปสมการเสียใหม่ (Xj หรือ f หรือทั้งสองประการ)

อย่างไรก็ตาม ถ้าไม่มี replication (ซึ่งมักเป็นเรื่องปกติ) เราก็ไม่ต้องทดสอบ lack of fit

4.6.2 การเลือก best subset model

เนื่องจากในปัจจุบันเราสามารถนำคอมพิวเตอร์เข้ามาช่วยงานวิเคราะห์ข้อมูลกันได้ง่าย เนื่องจากราคาถูกและมีโปรแกรมสำเร็จรูปเช่น SPSS, TSP, SAS ให้เลือกใช้ตามความถนัด เราจึงสามารถวิเคราะห์สมการถดถอยได้ด้วยวิธี all possible regression ซึ่งในสมัยก่อนถือว่าไม่เหมาะสมเพราะสิ้นเปลืองกำลังแรงงานมากได้อย่างมีประสิทธิภาพ ซึ่งวิธีนี้ในปัจจุบันนับได้ว่าเป็นวิธีที่ดีที่สุดได้ (อ่านตอน 4.5.2) เพราะนักวิจัยสามารถคัดเลือกเอาสมการใดสมการหนึ่งในบรรดาสมการถดถอยที่มีได้ถึง

$$\sum_{i=1}^{k-1} \binom{k-1}{i} \quad \text{หรือ} \quad 2^{(k-1)} - 1 \text{ สมการ}$$

มาใช้ได้ตามที่พิจารณาเห็นว่าเหมาะสม คือเรามีสิทธิ์เลือก มีสิทธิ์ใช้วิจารณญาณ เพื่อเลือกใช้สมการที่เราเห็นว่าเหมาะสมจากสมการต่างๆ ที่มีให้เลือกได้ถึง $2^{(k-1)} - 1$ สมการ ($k-1$ = จำนวนตัวแปรอิสระ)

เกณฑ์การเลือกอาจมีได้หลายแนวเช่น (1) R^2 สูงและ s^2 ต่ำ (2) ไม่มีปัญหา autocorrelation (3) ไม่มีปัญหา heteroscedastic disturbance (4) ไม่มีปัญหา error in veriables (5) ไม่มีปัญหา nonnormality (6) ไม่มีปัญหา multicolinearity (7) มีอำนาจพยากรณ์สูง (8) ขนาดของ $\hat{\beta}$ ไม่ขัดกับ prior information (9) ตัวแปรอิสระ X_j ต้องไม่ขัดกันทฤษฎีหรือศาสตร์เฉพาะทาง (10) อื่นๆ

เกณฑ์คัดเลือกทั้ง 9 ประการข้างต้นนี้เราต้องใช้ทุกประการพร้อมกันจะทิ้งประการใดไปเสียบ้างจะไม่เกิดผลดี ดังนี้การพิจารณาจึงเสียเวลามาก เช่นหากมีตัวแปรอิสระ 7 ตัวเราจะต้องพิจารณาคัดเลือก best subset ถึง $2^7 - 1 = 127$ สมการด้วยเกณฑ์ 9 ประการ เรื่องนี้นับว่าเสียเวลามากเราน่าจะมีวิธีกรองเอาสมการที่เข้าข่ายเพียงบางส่วนมาคัดเลือกภายใต้เกณฑ์ทั้ง 9 ประการ

วิธีกรองหรือคัดเลือก best subset มีหลายวิธี วิธีที่จะแนะนำไว้ในที่นี้คือ Mallow's Cp plot (เสนอโดย Mallows, C.L. ในปี 1973)

Cp plot มีวิธีปฏิบัติดังนี้

<u>ขั้นที่ 1</u> วิเคราะห์สมการถดถอยทั้งสิ้น $2^{k-1}-1$ สมการ แต่ละสมการให้คำนวณหา SSE (residual sum of squares) และ s^2 รวมทั้งตัวสถิติที่จำเป็นสำหรับการคัดเลือก ถ้าเป็นไปได้ให้ใช้ทั้งหมดคือทั้ง 9 ประการหรือมากกว่า

<u>ขั้นที่ 2</u> คำนวณหาค่า Cp

$$cp = \frac{\text{SSE ของ subset model ซึ่งมีพารามิเตอร์ p ตัว}}{s^2 \text{ ของสมการที่มีตัวแปรครบ k-1 ตัว}} - (n-2p)$$

ซึ่งเราจะได้ค่า cp รวมทั้งสิ้น $2^{k-1}-1$ ค่าเท่ากับจำนวน all possible regression model ทั้งนี้ p คือจำนวนพารามิเตอร์ทั้งหมดใน subset model (รวม intercept term)

<u>ขั้นที่ 3</u> นำคู่ลำดับ (p, Cp) ทั้งสิ้น $2^{k-1}-1$ คู่ลำดับไปพล็อตในระนาบที่มี p เป็นแกนนอนและ Cp เป็นแกนตั้ง

<u>ขั้นที่ 4</u> เลือกสมการที่สอดคล้องกับคู่ลำดับ (Cp, Cp) ที่วางอยู่บนเส้นตรง ที่เราลากจากจุดกำเนิดและทำมุม 45 กับแกนมาใช้ สมการดังกล่าวนี้เราถือได้ว่าเป็น best subset model แต่ก็ควรนำไปคัดเลือกตามเกณฑ์ 9 ประการ ข้างต้นด้วย

เรื่องนี้ขอให้นักศึกษาทดลองทำ Cp-plot โดยอาศัยข้อมูลจาก all possible regression solution for the Hald data ที่ปรากฏในหน้าถัด ๆ ไป ตามตัวอย่างนั้น จะพบว่ามีตัวแปรอิสระ 4 ตัวจึงมี subset model ได้ทั้งสิ้น $2^4-1 = 15$ สมการ

4.7 Multivariate Regression

ในทางปฏิบัตินั้นเราอาจพบว่าตัวแปรอิสระชุดเดิมคือ X's ซึ่งก็คือ $X_2, X_3 ..., X_k$ นั้น สามารถควบคุมพฤติกรรมของตัวแปรตามได้หลายตัวหรือหลายเรื่อง สิ่งนี้ขึ้นอยู่กับผู้วิจัยเองคือ ผู้วิจัยมี target variable Y หลายตัว หรือต้องการศึกษาและพยากรณ์หลายเรื่อง (เรื่องหนึ่งคือตัวแปร Y ตัวหนึ่ง) โดยเชื่อว่าตัวแปรที่สนใจนั้นๆ ถูกควบคุมอยู่ด้วยตัวแปรอิสระชุดเดียวกัน เช่น ผู้วิจัยเชื่อว่าดัชนีราคาสินค้าที่จำเป็น 7 หมวดคือ อาหารและเครื่องดื่ม (Y_1) เครื่องนุ่งห่ม (Y_2) การรักษาพยาบาล (Y_3) การเคหะ (Y_4) การขนส่ง (Y_5) การศึกษา การอ่านและการบันเทิง (Y_6) และเครื่องดื่มที่มีอัลกอฮอล์ (Y_7) ล้วนแต่มีค่าเคลื่อนไหวขึ้นลงตามราคาน้ำมันดิบหรือผลิตภัณฑ์น้ำมัน 4 รายการคือ น้ำมันเบนซินธรรมดา (X_2) น้ำมันเบนซินพิเศษ (X_3) น้ำมันดีเซลหมุนเร็ว (X_4) และก๊าซหุงต้ม (X_5) เรื่องนี้ผู้วิจัยมีสิทธิเลือกวิเคราะห์สมการถดถอยได้ 2 แนวคือ

1. วิเคราะห์แต่ละ Y กับ X_2, X_3, X_4, X_5 คือวิเคราะห์สมการถดถอย

$$Y_{ji} = \beta_{j1} + \beta_{j2} X_{2i} + \beta_{j3} X_{3i} \beta_{j4} X_{4i} + \beta_{j5} X_{5i} + u_{ji}$$

$$; \ i = 1, 2, ..., n \quad ; \quad j = 1, 2,..., 7$$

ซึ่งจะได้สมการถดถอยรวม 7 สมการ แปลว่าเราต้องวิเคราะห์สมการถดถอย 7 ครั้ง วิธีนี้คือวิธีปกติที่ได้ศึกษาผ่านมาแล้ว เรียกว่า univariate regression analysis

2. วิเคราะห์ Y ทุกตัวพร้อมกันกับ X_2,X_3,X_4,X_5 ซึ่งจะได้สมการถดถอยรวม 7 สมการ แต่เราวิเคราะห์เพียงคราวเดียว วิธีนี้เรียกว่า multivariate regression analysis แนวทางทั้งสองให้ผลลัพธ์ตรงกัน

ในที่นี้จะกล่าวถึง Multivariate Regression โดยสังเขปพอให้เห็นลักษณะทั่วไปเท่านั้น หากนักศึกษาสนใจอาจติดตามศึกษาได้จากเอกสารอ้างอิง และอาจศึกษาได้จากกระบวนวิชา Multivariate Statistical Analysis.

ให้ $Y_1, Y_2,..., Y_m$ เป็นตัวแปรตามที่ต่างก็ถูกควบคุมหรือผูกพันอยู่กับตัวแปรอิสระชุดเดียวกันคือ $X_1, X_2,...,X_k$ โดยที่ $X_1 = [1,1, ..., 1]'$

เราจึงมีสมการถดถอย Y = f(X's) รวม m สมการคือ

$$Y_{(i)} = X\beta_{(i)} + U_{(i)};\ i = 1, 2, ..., m\ ;\ t = 1, 2, ..., n$$

ในที่นี้ (i) หมายถึงลำดับที่ของสมการ

เมื่อนำสมการทั้ง m สมการมาเสนอร่วมกันจะปรากฏลักษณะดังนี้

$$[Y_{(i)} \mid Y_{(2)} \mid ... \mid Y_{(m)}] = X\,[\beta_{(1)} \mid \beta_{(2)} \mid ... \mid \beta_{(m)}] + [U_{(1)} \mid U_{(2)} \mid ... \mid U_{(m)}]$$

หรือ $$\underset{(nxm)}{Y} = \underset{(nxk)}{X}\ \underset{(kxm)}{\beta} + \underset{(nxm)}{U}$$

$$\begin{bmatrix} Y_{11} & Y_{21} & \cdots & Y_{m1} \\ Y_{12} & Y_{22} & \cdots & Y_{m2} \\ Y_{13} & Y_{23} & \cdots & Y_{m3} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1n} & Y_{2n} & \cdots & Y_{mn} \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ 1 & X_{23} & X_{33} & \cdots & X_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} & \cdots & \beta_{m1} \\ \beta_{12} & \beta_{22} & & \beta_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{1k} & \beta_{2k} & \cdots & \beta_{mk} \end{bmatrix} + U$$

$$U = \begin{bmatrix} U_{11} & U_{21} & \cdots & U_{m1} \\ U_{12} & U_{22} & \cdots & U_{m2} \\ U_{13} & U_{23} & \cdots & U_{m3} \\ \vdots & \vdots & \vdots & \vdots \\ U_{1n} & U_{2n} & \cdots & U_{mn} \end{bmatrix}$$

โดยที่ $E(U(i)) = \underset{(nx1)}{0}$, $V(U(i)) = Q_i^2 I_n$, $\text{cov}\,(U(i), U(l)) = \sigma_{il} I_n$

; i,l = 1, 2, ... m

หมายความว่าเมื่อพิจารณาเฉพาะสมการลำดับที่ (i) ข้อตกลงก็ยังคงเป็นเช่นเดิม แต่ถ้าพิจารณาระหว่างสมการลำดับที่ (i) กับลำดับที่ (l) แล้ว residual ของสมการคู่นี้อาจสัมพันธ์กันได้ เมื่อพิจารณาที่สมการลำดับ (i) ใดๆ คือ

$$\underset{(nx1)}{Y_{(i)}} = \underset{(nxk)}{X}\ \underset{(kx1)}{\beta(i)} + \underset{(nx1)}{U(i)}\ \ ;\ i = 1, 2, ..., m$$

จะพบว่า $\hat{\beta}_{(i)} = (X'X)^{-1}X'Y_{(i)}$; i = 1, 2, ... , m

เรียก $\hat{\beta}_{(i)}$ ว่า univariate OLS ซึ่งได้มาจากการ minimize $(Y(i)-X\hat{\beta}_{(i)})'(Y(i)-X\hat{\beta}_{(i)})$

หรือ $e'_{(i)} e_{(i)}$ ตามปกติ

ถ้านำผลลัพธ์คือ $\hat{\beta}_{(i)}$ มาเรียงต่อกันจะได้

$$\hat{\beta} = [\hat{\beta}_{(1)} \mid \hat{\beta}_{(2)} \mid \cdots \mid \hat{\beta}_{(m)}]$$

$$= (X'X)^{-1}X' [Y_{(1)} \mid Y_{(2)} \mid \cdots \mid Y_{(m)}]$$

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

และ $\hat{Y} = X\hat{\beta}$

นอกจากนี้เรายังสามารถพิสูจน์ได้ว่า

1) $E(\hat{\beta}_{(i)}) = \beta_{(i)}$

2) $\text{Cov}(\hat{\beta}_{(i)}, \hat{\beta}_{(l)}) = E[\hat{\beta}_{(i)} - \beta_{(i)}] [\hat{\beta}_{(l)} - \beta_{(l)}]'$

$= E[(X'X)^{-1}X'U_{(i)}] [(X'X)^{-1}X'U_{(l)}]'$

$= (X'X)^{-1}X'E(U_{(i)}U'_{(l)}) X(X'X)^{-1}$

$= \sigma_{il}(X'X)^{-1}$; i, l = 1, 2, ..., m

$E(e'_{(i)}e_{(l)}) = E[U'_{(i)}\{I_n - X(X'X)^{-1}X'\}U(l)]$

$= \text{tr}[\{In - X(X'X)^{-1}X'\}\sigma_{il} I_n]$

$= \sigma^2_{il}(n-k)$

$s^2_{il} = \hat{\sigma}^2_{il} = \frac{1}{n-k} e'^{(i)}e^{(l)}$; i, l = 1,2,...,m

$= \frac{1}{n-k}(Y_{(i)} - X\hat{\beta}_{(i)})'(Y_{(l)} - X\hat{\beta}_{(l)})$

ตัวอย่างเช่น ผลจากการบันทึกข้อมูลปรากฏดังนี้

X	0	1	2	3	4
$Y_{(1)}$	1	4	3	8	9
$Y_{(2)}$	-1	-1	2	3	2

จงวิเคราะห์สมการถดถอย

$$Y_{1i} = \beta_{11} + \beta_{12}X_i + U_{1i}$$
$$Y_{2i} = \beta_{21} + \beta_{22}X_i + U_{2i} \quad , i = 1, 2, ..., 5$$

วิธีทำ

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \beta = [\beta_{(1)} | \beta_{(2)}] = \begin{bmatrix} \beta 11 & \beta 21 \\ \beta 12 & \beta 22 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = [\hat{\beta}_{(1)} | \hat{\beta}_{(2)}]$$

สมการถดถอยทั้งสองที่ต้องการคือ $Y_{(1)} = 1+2X$ และ $Y_{(2)} = -1+X$

$$e = [e_{(1)} | e_{(2)}] = Y - \hat{Y} = Y - X\hat{\beta} = \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \\ -2 & 1 \\ 1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$s_{11} \text{ หรือ } s^2_1 = \frac{1}{n-2} e'_{(1)}e_{(1)} = \frac{6}{3} = 2$$

$$s^2_2 = \frac{1}{n-2} e'_{(2)}e_{(2)} = \frac{4}{3} = 1.33$$

$$s_{12} = \frac{1}{n-2} e'_{(1)}e_{(2)} = -\frac{2}{3} = -.66$$

ขอให้สังเกตว่าแทนที่เราจะวิเคราะห์แบบ multivariate regression เรากลับทำการวิเคราะห์แบบ univariate regression 2 ครั้ง ๆ ละสมการก็ได้ผลลัพธ์ตรงกัน เหตุนี้ multivariate regression จึงเป็นเรื่องฟุ่มเฟือยอยู่สักหน่อย

STEP-WISE SOLUTION FOR THE HALD DATA

Original and/ or Transformed Data

	X_1	X_2	X_3	X_4	X_5
1	7.00000000	26.00000000	6.00000000	60.00000000	78.50000000
2	1.00000000	29.00000000	15.00000000	52.00000000	74.30000000
3	11.00000000	56.00000000	8.00000000	20.00000000	104.30000000
4	11.00000000	31.00000000	8.00000000	47.00000000	87.60000000
5	7.00000000	52.00000000	6.00000000	33.00000000	95.90000000
6	11.00000000	55.00000000	9.00000000	22.00000000	109.20000000
7	3.00000000	71.00000000	17.00000000	6.00000000	102.70000000
8	1.00000000	31.00000000	22.00000000	44.00000000	72.50000000
9	2.00000000	54.00000000	18.00000000	22.00000000	93.10000000
10	21.00000000	47.00000000	4.00000000	26.00000000	115.90000000
11	1.00000000	40.00000000	23.00000000	34.00000000	83.80000000
12	11.00000000	66.00000000	9.00000000	12.00000000	113.30000000
13	10.00000000	68.00000000	8.00000000	12.00000000	109.40000000

Means of Transformed Variables

1	7.46153830	48.15384500	11.76923000	29.99999900	95.42307500

Std. Deviations of Transformed Variables

1	5.88239440	15.56087900	6.40512590	16.73817800	15.04372400

Correlation Matrix

1	.99999991	.22857948	-.82413372	-.24544512	.73071745
2	.22857948	1.00000010	-.13924238	-.97295516	.81625268
3	-.82413372	-.13924238	.99999991	.02953701	-.53467065
4	-.24544512	-.97295516	.02953701	1.00000010	-.82130513
5	.73071745	.81625268	-.53467065	-.82130513	.99999999

Control Information

No. of observations	13
F level for entering a variable	3.28
F level for deleting a variable	3.28
Response variable is no.	5
Risk level for B conf. interval	5%

Step No. 1

Variable entering	4
Sequential F-test	22.7985280
Pereent variation explained R-SQ	67.4542100
Standard error of Y	8.9639014
Mean of the response	95.4230750
Std. error as a % of mean response	9.394%
Degrees of freedom	11
Determiaant value	1.0000001

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	1	1831.8968000	1831.8968000	22.7985300
Residual	11	883.8668200	80.3515290	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.7381620	-.3978962 -1.0784277	.1545960	22.7985270

Constant Term in Prediction Equation 117.5679300

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Squere of Partials
1	.91541
2	.01696
3	.80117
5	1.00000

Step No. 2

Variable entering	1
Sequential F-test	108.2240500
Percent variation explained R-SQ	97.2471100
Standard error of Y	2.7342642
Mean of the response	95.4230750
Std. error as a % of mean response	2.865%
Degrees of freedom	10
Determinant value	.9397567

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	2	2641.0015000	1320.5007000	176.6272400
Residual	10	74.7620080	7.4762008	

B Coeffieients and Confidanee Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.6139538	-.5055738 -.7223338	.0486445	159.2954900
1	7.4615383	1.4399582	1.7483502 1.1315662	.1384165	108.2240500

Constant Term in Prediction Equation 103.0973800

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
2	.35833
3	.32003
5	1.00000

Step No. 3

Variable entering	2
Sequential F-test	5.0258747
Percent variation explained R-SQ	98.2335500
Standard error of Y	2.3087426
Mean of the response	95.4230750
Std. error as a % of mean response	2.419%
Degrees of freedom	9
Determinant value	.0500394

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	3	2667.7908000	889.2636000	166.8320500
Residual	9	47.9726310	5.3302923	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.2365401	.1554367 -.6285170	.1732877	1.8632619
1	7.4615383	1.4519379	1.7165861 1.1872897	.1169975	154.0079500
2	48.1538450	.4161100	.8359608 -.0037408	.1856104	5.0258730

Constant Term in Prediction Equation 71.6482910

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.00227
5	1.00000

Step No. 4

Variable leaving is	4
Sequential F-test	1.8632611
Pereent variation explained R-SQ	97.8678500
Standard error of Y	2.4063325
Mean of the response	95.4230750
Std. error as a % of mean response	2.522%
Degrees of freedom	10
Determinant value	.9477514

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	2	2657.8593000	1328.9296000	229.5042500
Residual	10	57.9043570	5.7904357	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
1	7.4615383	1.4683057	1.7385638 1.1980476	.1213008	146.5229500
2	48.1538450	.6622507	.7644149 .5600864	.0458547	208.5821200

Constant Term in Prediction Equation 52.5773400

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.16914
4	.17152
5	1.00000

Residual Analysis

Obs. No.	Observed Y	Predicted Y	Residual	Normal Deviate
1	78.5000000	80.0739960	-1.5739960	-.6541058
2	74.3000000	73.2509140	1.0490860	.4359688
3	104.3000000	105.8147300	-1.5147300	-.6294766
4	87.6000000	89.2584720	-1.6584720	-.6892115
5	95.9000000	97.2925130	-1.3925130	-.5786869
6	109.2000000	105.1524800	4.0475200	1.6820285
7	102.7000000	104.0020500	-1.3020500	-.5410931
8	72.5000000	74.5754150	-2.0754150	-.8624806
9	93.1000000	91.2754870	1.8245130	.7582132
10	115.9000000	114.5375400	1.3624600	.5661977
11	83.8000000	80.5356710	3.2643290	1.3565577
12	113.3000000	112.4372400	.8627600	.3585373
13	109.4000000	112.2934400	-2.8934400	-1.2024273

จากข้อมูลชุดเดียวกันนี้ (Hald's data) เมื่อสั่งวิ่งโปรแกรม SPSS PC + ตาม command line ต่อไปนี้ปรากฏผลลัพธ์ดังนี้

```
DATA LIST FILE = "B : DATA.DAT"
    /X1 3-14 X2 17-28 X3 31-41 X4 45-56 X5 59-70.
SET PRINT ON/MORE OFF.
REGRESSION DESCRIPTIVES = DEFAULTS
    /VARIABLES = X1 TO X5
    /STATISTICS = R ANOVA CHA BCOV COEFF OUTS ZPP CI SES
      TOL LINE HISTORY END
    /DEPENDENT = X5
    /METHOD = FORWARD
    /RESIDUAL = HISTOGRAM OUTLIERS (*DRESID) NORMPROB DURBIN
    /CASEWISE = ALL DEPENDENT PRED RESID SRESID DRESID MAHAL COOK
    /DEPENDENT = X5
    /METHOD = BACKWARD
    /RESIDUAL = HISTOGRAM OUTLIERS (*DRESID) NORMPROB DURBIN
    /CASEWISE = ALL DEPENDENT PRED RESID SRESID DRESID MAHAL COOK
    /DEPENDENT = X5
    /METHOD = STEPWISE
    /RESIDUAL = HISTOGRAM OUTLIERS (*DRESID) NORMPROB DURBIN
    /CASEWISE = ALL DEPENDENT PRED RESID SRESID DRESID MAHAL COOK
    /DEPENDENT = X5
    /METHOD = ENTER X1 TO X5
    /RESIDUAL = HISTOGRAM OUTLIERS (*DRESID) NORMPROB DURBIN
    /CASEWISE = ALL DEPENDENT PRED RESID SRESID DRESID MAHAL COOK
    /SCATTERPLOT = (X1, X5)(X2, X5)(X3, X5)(X4, X5)(*PRED, *RESID).
```

ข้อมูล (Hald's data) ในแฟ้ม DATA. DAT ปรากฏดังนี้

```
 7.00000000   26.00000000    6.00000000   60.00000000   78.50000000
 1.00000000   29.00000000   15.00000000   52.00000000   74.30000000
11.00000000   56.00000000    8.00000000   20.00000000  104.30000000
11.00000000   31.00000000    8.00000000   47.00000000   87.60000000
 7.00000000   52.00000000    6.00000000   33.00000000   95.90000000
11.00000000   55.00000000    9.00000000   22.00000000  109.20000000
 3.00000000   71.00000000   17.00000000    6.00000000  102.70000000
 1.00000000   31.00000000   22.00000000   44.00000000   72.50000000
 2.00000000   54.00000000   18.00000000   22.00000000   93.10000000
21.00000000   47.00000000    4.00000000   26.00000000  115.90000000
 1.00000000   40.00000000   23.00000000   34.00000000   83.80000000
11.00000000   66.00000000    9.00000000   12.00000000  113.30000000
10.00000000   68.00000000    8.00000000   12.00000000  109.40000000
```

SPSS/PC+ The Statistical Package for IBM PC
The raw data or transformation pass is proceeding
13 cases are written to the uncompressed active file.

Page 2 SPSS/PC+

* * * * MULTIPLE REGRESSION * * * *

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
X1	7.462	5.882	
X2	48.154	15.561	
X3	11.769	6.405	
X4	30.000	16.738	
X5	95.423	15.044	

N of Cases = 13

Correlation :

	X1	X2	X3	X4	X5
X1	1.000	.229	−.824	−.245	.731
X2	.229	1.000	−.139	−.973	.816
X3	−.824	−.139	1.000	.030	−.535
X4	−.245	−.973	.030	1.000	−.821
X5	.731	.816	−.535	−.821	1.000

โปรแกรมเริ่มทำคำสั่ง / DESCRIPTIVES แสดงมัชฌิมเลขคณิตและส่วนเบี่ยงเบนมาตรฐานของตัวแปร X1 ถึง X5 ไม่พิมพ์ label คือชื่อเต็มของตัวแปร เพราะเราไม่ได้สั่งคำสั่ง VARIABLE LABELS และแสดง correlation table ด้วย ถ้าพิจารณาจากแถวที่ 5 หรือสดมภ์ที่ 5 ของ correlation table จะพบสหสัมพันธ์ระหว่าง Y กับ X's เป็น $r_{1y} = .731$ $r_{2y} = .816$ $r_{3y} = -.535$ และ $r_{4y} = -.821$ ตามลำดับ

การวิเคราะห์เริ่มด้วยการหา best fitted model ตามวิธี FS โดยนำเอา summary statistics มาแสดงก่อน เพราะเราสั่งคำสั่งว่า / STATISTICS = HISTORY เอาไว้ เพื่อจะได้ทราบผลการทำงานทั้งหมดไว้เป็นการล่วงหน้าดังนี้

Page 3 SPSS/PC+

* * * * MULTIPLE REGRESSION * * * *

Page 4 SPSS/PC+

* * * * MULTIPLE REGRESSION * * * *

Equation Number 1 Dependent Variable... X5

Beginning Block Number 1. Method : Forward

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1	.8213	.6745	22.799	.001	In :	X4	−.8213
2	.9861	.9725	176.627	.000	In :	X1	.5631

summary table แสดงให้ทราบว่า เมื่อเราสั่ง / METHOD = FORWARD โปรแกรมจะเริ่มด้วยการนำ X4 เข้าสู่สมการ เป็น Y = f(X4) โดยรายงานว่า X4 และ X5 (คือ Y) มี multiple correlation เท่ากับ .8213 (ความจริงก็คือ R_{4y} = .8213) $R^2 = (.8213)^2 = .6745$ F_c = 22.799 Sig F = .001 < .05 แสดงว่า $H_0 : \beta_4 = 0$ มีนัยสำคัญ (คือเราปฏิเสธ H_0) คำว่า In หมายถึงตัวแปรภายในสมการ In : X4 แปลว่า X4 เป็นตัวแปรในสมการ คือ Y = f(X4) BetaIn ก็คือค่าประมาณของ ρ (อย่าลืมว่า $\hat{\rho}_4 = \frac{s_4}{s_y}\hat{\beta}_4$) เป็น สปส ของสมการถดถอยในกรณี standardized model เหตุที่เสนอเป็น Beta ไม่เสนอเป็น B (คือ $\hat{\beta}$) คงประสงค์จะให้เราเปรียบเทียบปริมาณอิทธิพลของตัวแปรอิสระในสมการ

ในลำดับชั้น (step) ที่ 2 โปรแกรมจะนำ X1 เข้าสู่สมการ ได้สมการเป็น Y = f(X4, X1) การมี X1 เพิ่มเข้ามาทำให้ multiple correlation เปลี่ยนเป็น .9861 หรือ $R^2 = (.9861)^2 = .9725$ F_c = 176.627 Sig F = .000 < .05 แสดงว่าสมมติฐาน $H_0 : \beta_4 = \beta_1 = 0$ มีนัยสำคัญ (หรือปฏิเสธ H_0) และ $\hat{\rho}_1$ = .5631

การวิเคราะห์สมการถดถอยตามวิธี FORWARD selection ดำเนินการดังนี้

Page 5 SPSS/PC+

* * * * M U L T I P L E R E G R E S S I O N * * * *

Equation Number 1 Dependent Variable... X5

Variable (s) Entered on Step Number
2... X1

Multiple R	.98614		
R Square	.97247	R Square Change	.29793
Adjusted R Square	.96697	F Change	108.22391
Standard Error	2.73427	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	2641.00096	1320.50048
Residual	10	74.76211	7.47621

F = 176.62696 Signif F = .0000

อ่านว่า ใน step ที่ 2 มีตัวแปรอิสระเพิ่มเข้ามา คือ X1 สถิติต่าง ๆ ข้างบนเป็นผลจากการสั่งว่า

/STATISTICS = R ANOVA CHA BCOV COEFF OUTS ZPP CI SES TOLF LINE
HISTORY END

keyword R ทำให้ได้สถิติ R, R^2 และ R^2adj

keyword ANOVA ทำให้ได้ตาราง Analysis of variance

keyword CHA ทำให้ทราบความเปลี่ยนแปลงในค่าของ R^2 F_c และ Sig F ว่าเปลี่ยนแปลงไปจาก step ก่อนหน้านี้มากน้อยเพียงใด จากผลลัพธ์ข้างบน

R^2 change = .29793 แสดงว่าใน step ที่ 2 ซึ่งเพิ่ม X1 ลงในสมการถดถอย ทำให้ R^2 เพิ่มสูงขึ้น .29793 ค่า F_c เพิ่มขึ้น 108:22391 Signif F change = .000 แสดงว่าพื้นที่ใต้โค้งด้านขวามือของ F_c ไม่เพิ่มขึ้น

ใต้ตาราง Analysis of variance รายงานว่า F = 176.62696 Signif F = .0000 แสดงว่า $H_0 : \beta_4 = \beta_1 = 0$ มีนัยสำคัญ (เพราะ Signif F = .0000 < .05)

keyword BCOV ทำให้ทราบ variance – covariance matrix ของเวคเตอร์

$\hat{\beta}' = (\hat{\beta}_4, \hat{\beta}_1)$ คือ $\hat{V}(\hat{\beta}) = s^2(X'X)^{-1}$ ดังนี้

Page 6 SPSS/PC+

**** MULTIPLE REGRESSION ****

Equation Number 1 Dependent Variable... X5

Var-Covar Matrix of Regression Coefficients (B)

Below Diagonal : Covariance Above : Correlation

	X4	X1
X4	.00237	.24545
X1	.00165	.01916

keyword COEFF ทำให้ได้ค่า B ในที่นี้คือ $\hat{\beta}_4$ และ $\hat{\beta}_1$ และ $\hat{\beta}_0$ (คือ constant) ขณะที่ keyword CI ให้ช่วงเชื่อมั่น 95% ของ B (เรียกว่า unstandardize regression coefficient) COEFF ยังให้ค่าประมาณของ standardized regression ceeficient ซึ่งปรากฏในสดมภ์ชื่อ Beta คือ $\hat{\rho}_4$ และ $\hat{\rho}_1$ ดังตารางต่อไปนี้ ขอให้สังเกตว่า SE B (standard error of unstandardized regression coef) ซึ่งแสดงไว้ที่นี้ด้วยก็คือ รากที่สองของสมาชิกในแนวทแยงของ BCOV ข้างบน คือ $\sqrt{.00237} = .04864$ และ $\sqrt{.01916} = .13842$ และแสดงค่า t และ Sig T ด้วย

keyword ZPP แสดงค่าสหสัมพันธ์ 3 ตัว คือ zero order correlation ซึ่งก็คือ r_{jy} ในที่นี้คือ r_{4y} กับ r_{1y} Part correlation และ Partial correlation

part correlation คือ สหสัมพันธ์ระหว่าง Y กับ X_j ภายหลังจากที่เราได้กำจัดหรือผลักอิทธิพลของตัวแปรอิสระอื่น ๆ ออกจาก X_j แล้ว นั่นคือ

$$\text{part corr.} = \sqrt{R^2\text{change}} = \sqrt{R^2 - R^2_{(j)}}$$

โดยที่ $R^2_{(j)}$ คำนวณได้มาจากสมการ Y = f (X's ยกเว้น X_j) เช่นในกรณีนี้เรามี X_4 กับ X_1 เราได้ R^2 จากสมการ $Y = f(X_4, X_1)$ ได้ $R^2_{(4)}$ จากสมการ $Y = f(X_1)$ ได้ $R^2_{(1)}$ จากสมการ $Y = f(X_4)$ เราจึงได้ part corr ของ $X_1 = \sqrt{R^2 - R^2_{(1)}}$ part corr. ของ $X_4 = \sqrt{R^2 - R^2_{(4)}}$

ตัวแปรอิสระใดให้ค่า part correlation สูงกว่า แสดงว่าตัวแปรอิสระนั้นมีผลหรือให้ข้อสนเทศหรือมีอิทธิพลต่อ Y สูงกว่าและตัวแปรอื่นไม่อาจให้ข้อสนเทศเช่นนั้นแก่ Y ทดแทนกันได้

**** MULTIPLE REGRESSION ****

Equation Number 1 Dependent Variable... X5

Variables in the Equation

Variable	B	SE B	95% Confdnce Intrvl B		Beta
X4	−.61395	.04864	−.72234	−.50557	−.68311
X1	1.43996	.13842	1.13155	1.74837	.56305
(Constant)	103.09738	2.12398	98.36485	107.82991	

Variables in the Equation

Variable	SE Beta	Correl	Part Cor	Partial	Tolerance	T	Sig T
X4	.05412	−.82131	−.66221	−.97002	.93976	−12.621	.0000
X1	.05412	.73072	.54583	.95677	.93976	10.403	.0000
(Constant)						48.540	.0000

จากตารางข้างบน part corr ของ $X_1 = \sqrt{R^2 - R_{(1)}} = \sqrt{.97247 - .6745}$ $= \sqrt{.29793} = .5458$

ขณะเดียวกันเราพบว่า $(\text{part corr})^2$ ก็คือ R^2change ซึ่งบอกให้เราทราบแต่เพียงว่าถ้าผนวกตัวแปรนี้เข้ามาในสมการ จะทำให้ R^2 เพิ่มขึ้นเท่าไร ซึ่งไม่ได้แปลว่าจะทำให้ Σe^2 (unexplained Sum Square) ลดลง ซึ่งหากในสมการของเรารับตัวแปรสำคัญไว้ก่อน ซึ่ง R^2 จะใกล้ 1 อยู่แล้ว ตัวแปรที่ตามมาภายหลังก็ย่อมให้ค่า part corr. ได้เพียงเล็กน้อย คือ ใกล้ 0 โดยปริยาย คราวนี้ความมั่นใจของเราที่มีต่อ part corr. ว่าสามารถใช้ไปเพื่อเปรียบเทียบอิทธิพลของตัวแปรอิสระก็เริ่มหวั่นไหวเสียแล้ว ทางออกก็คือ แทนที่จะใช้ part corr. กลับให้ใช้ partial correlation แทน

partial correlation คือ สหสัมพันธ์ระหว่าง Y กับ X_j ภายหลังจากที่เราได้ผลักอิทธิพลของตัวแปรอื่นออกไปจากทั้ง X_j และ Y ซึ่งจะเป็นค่าสหสัมพันธ์แท้ ๆ ระหว่าง Y กับ X_j โดยไม่มีตัวแปรอื่นมาสอดแทรกหรือเป็นสื่อประสาน

การคำนวณหา partial corr. หาได้จากสูตร $R^2 = \frac{SSR}{SSE}$ ดังนี้คือ

$$(\text{partial corr ของ } X_j)^2 = \frac{R^2 - R^2_{(j)}}{1 - R^2_{(j)}} \; ; \; j = 2(1)k$$

$R^2 - R^2_{(j)}$ = อิทธิพลของเฉพาะ X_j โดยที่ $R^2_{(j)}$ และ R คำนวณจากสมการ

$$Y = f(X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$$

และ $Y = f(X_1, X_2, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_k)$ ตามลำดับ

ตามผลลัพธ์ข้างต้น พบว่า Pr4 = .97002 และ Pr1 = .95677 แสดงว่า X_4 ผูกพันกับ Y สูงกว่า X_1 (จะเห็นว่าไม่ขัดกันกับ Beta)

keyword SES แสดงค่าประมาณ standard error ของ Beta

keyword TOLF แสดงค่า tolerance เรานิยาม tolerance ดังนี้

toler = $1 - R_j^2$ โดยที่ R_j^2 คำนวณจากสมการ auxiliary regression

$$X_j = f(X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$$

auxiliary regression คือสมการแสดงความผูกพันระหว่างตัวแปรอิสระที่ j กับตัวแปรอิสระที่เหลือ R_j^2 จึงเป็นดัชนีวัดดีกรีของ multicolinear ถ้า R_j^2 มีค่าสูงแสดงว่าน่าจะเกิดปัญหา multicolinearity ขึ้น เพราะสาเหตุว่า X_j เกิดจากการประกอบกันเชิงเส้น (linear combination) ของ X ตัวอื่น ๆ R_j^2 มีค่าสูง $1 - R_j^2$ จะมีค่าต่ำ ดังนั้นสิ่งที่เราพิจารณาก็คือตัวแปรอิสระตัวที่ให้ค่า toler ต่ำมาก ๆ ถ้าต่ำมากแปลว่าตัวแปรนั้นควรถูกเพ่งเล็งว่าจะทำให้เกิดปัญหา multicolinearity คำถามก็คือ toler มีค่าเท่าไรจึงถือว่าต่ำมาก โปรแกรม SPSS ถือเป็นข้อกำหนด (default) ว่า toler ต่ำสุดต้องไม่ต่ำกว่า 0.01 ถ้าตัวแปรใดให้ค่า toler ต่ำกว่า .01 โปรแกรม SPSS จะตัดทิ้งหรือเลิกพิจารณาแล้วรายงานผลว่า limit reached (ดูคำสั่ง / CRITERIA) ความจริงแล้ว toterance ก็คือ VIF นั่นเอง

อนึ่ง ผลจาก keyword COEFF ทำให้โปรแกรมแสดงค่า t และ Sig t จากตาราง พบว่า $t_4 = -12.621$ Sig $t_4 = .0000 < .05$ $t_1 = 10.403$ Sig $t_1 = .0000 < .05$ และ $t_0 = 48.540$ Sig $t_0 = .0000 < .05$ แสดงว่าสมมติฐาน $H_0: \beta_4 = 0$ $H_0: \beta_1 = 0$ และ $H_0: \beta_0 = 0$ ล้วนแต่มีนัยสำคัญด้วยกันทุกสมมติฐาน keyword OUTS แสดงสถิติต่าง ๆ ของตัวแปรนอกสมการ โดยแสดงให้เราได้ทราบว่า หากตัวแปรภายนอกเหล่าถูกส่งเข้าสู่สมการใน step ต่อไปจะเกิดผลลัพธ์เช่นในตาราง ทำให้เตรียมตัวได้ถูกว่าควรทำอย่างไรต่อไป

Page 8 SPSS/PC+

* * * * MULTIPLE REGRESSION * * * *

Equation Number 1 Dependent Variable... X5

Variables not in the Equation

Variable	Beta In	Partial	Tolerance	Min Toler	T	Sig T
X2	.43041	.59861	.05325	.05280	2.242	.0517
X3	−.17458	−.56571	.28905	.27187	− 2.058	.0697

End Block Number 1 PIN = .050 Limits reached.

เช่น จากผลลัพธ์ข้างบนทำให้เราทราบว่า ใน step ต่อไปถ้าเราเพิ่ม X2 เข้าสู่สมการจะทำให้ได้ $\beta_2 = .43041$ Pr2 = .59861 หมายความว่าถ้าผลักอิทธิพลของ X4 และ X1 ซึ่งอยู่ในสมการแล้วออกจากทั้ง Y และจาก X2 จะพบว่าความสัมพันธ์ระหว่าง X2 กับ Y เท่ากับ .59861 X2 ให้ค่า toler เท่ากับ .05325 กล่าวคือ $1 - R_2^2 = .05325$ โดยที่ R_2^2 คำนวณ

จากสมการ $X_2 = f(X_4, X_1)$ และถ้าเติม X_2 ลงไปในสมการจะพบว่า $t_2 = 2.242$ และ Sig t = .0517 > .05 ซึ่งแสดงว่าสมมติฐาน $H_0 : \beta_2 = 0$ ไม่มีนัยสำคัญ

summary table ปรากฏอีกครั้งหนึ่งดังนี้

Page 9 . SPSS/PC+

**** MULTIPLE REGRESSION ****

Equation Number 1 Dependent Variable... X5

Summary table

Step	MultR	Rsq	F(Eqn)	SigF	Variable	BetaIn
1	.8213	.6745	22.799	.001	In: X4	−.8213
2	.9861	.9725	176.627	.000	In: X1	.5631

สรุปผลการวิเคราะห์สมการถดถอยตามวิธี forward selection ที่ผ่านมาปรากฏดังนี้

$Y = 103.09738 - 0.61395X_4 + 1.43996X_1$

(2.12398) (0.04864) (0.13842)

(48.540*) (- 12.621") (10.403")

; $R^2adj = .96697$, $s^2 = 7.47621$, DW = 1.78765, F = 176.627"

สมการมีลักษณะที่ตรวจด้วย 1st order test แล้วน่าพอใจ เมื่อตรวจดูที่สดมภ์ชื่อ Beta เห็นได้ว่า X_4 มีอิทธิพลต่อ Y สูงกว่า X_1 เล็กน้อย

ผลจากคำสั่ง / RESIDUALS ปรากฏผลลัพธ์ดังนี้

Page 10 SPSS/PC+

**** MULTIPLE REGRESSION ****

Equation' Number 1 Dependent Variable... X5

Residuals Statistics

	Min	Max	Mean	Std Dev	N
'PRED	72.6117	117.3737	95.4231	14.x3.52	13
*ZPRED	- 1.5376	1.4796	.0000	I.0000	13
*SEPRED	.7719	1.9864	1.2748	.3294	13
*ADJPRED	71.9737	119.0207	95.5452	15.1195	13
*RESID	- 5.0234	3.7701	−.0000	2.4960	13
*ZRESID	- 1.8372	1.3788	- .0000	.9129	13
*SRESID	-2.0617	1.4671	−.0184	I.0241	13
*DRESID	-6.3264	4.26X4	−.1221	3.1748	13
"SDRESID	-2.5796	I.5712	−.0489	1.1256	13
*MAHAL	.0334	5.4100	I.8462	I.4659	13
"COOK D	.0002	.3675	.0929	.1071	13
*LEVER	.0028	.4508	.1538	.1222	13

Total Cases = 13

* * * * MULTIPLE REGRESSION * * * *

Equation Number 1 Dependent Variable... X5

Durbin-Watson Test = 1.78765

residual statistics เป็น temporary variable ดังนั้นจึงมีเครื่องหมาย *' กำกับข้างหน้าตารางข้างบนแสดงสถิติต่าง ๆ ของค่าพยากรณ์และ residual รวม 4 รายการ คือค่าต่ำสุด ค่าสูงสุด มัชฌิมเลขคณิต และส่วนเบี่ยงเบนมาตรฐาน

PRED หมายถึงค่าพยากรณ์ของ Y_i (predicted Y_i) ซึ่งก็คือ $\hat{Y}_i$

ZPRED หมายถึง standardized predicted value

เรื่องนี้มีคำอธิบายทางทฤษฎีดังนี้

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X(X'X)^{-1}X'Y \\ &= [I - X(X'X)^{-1}X']Y \\ &= [I - H]Y \end{aligned}$$ โดยที่ $H = X(X'X)^{-1}X'$ เรียกว่า hat matrix

จะพบว่า $V(e) = E[(e - E(e))(e - E(e))']$

$$\begin{aligned} &= E(ee') \\ &= E[(I - H)YY'(I - H)'] \\ &= (I - H)E(YY')(I - H)' \\ &= \sigma^2(I - H) \end{aligned}$$

และพบว่า $V(e_i) = \sigma^2(1 - h_{ii})$ เมื่อ h_{ii} คือสมาชิกที่ (i, i) ของ H แต่เราสามารถหา h_{ii} โดยตรงโดยไม่ต้องเรียกจาก H ได้ดังนี้

เนื่องจาก $e_i = Y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_{k-1} X_{(k-1)i}$

$= Y_i - X_i'\hat{\beta}$ โดยที่ X_i' คือแถวที่ i ของเมตริกซ์ X

ดังนั้น $V(e_i) = V(Y_i) - V(X_i'\hat{\beta}) = V(Y_i) - V(\hat{Y}_i)$

ซึ่งจะพบว่า

$$\begin{aligned} V(Y_i) &= E[(Y_i - E(Y_i))(Y_i - E(Y_i))] \\ &= E[(Y_i - X_i'\beta)(Y_i - X_i'\beta)] \\ &= E[(X_i'\beta + u_i - X_i\beta)(X_i'\beta + u_i - X_i'\beta)] \\ &= E(u_i^2) = V(u_i) \\ &= \sigma^2 \end{aligned}$$

$$
\begin{aligned}
V(\hat{Y}_i) = V(X_i'\hat{\beta}) &= E[(X_i'\hat{\beta} - E(X_i'\hat{\beta}))(X_i'\hat{\beta} - E(X_i'\hat{\beta}))] \\
&= E[(X_i'\hat{\beta} - E(X_i'\hat{\beta}))(X_i'\hat{\beta} - E(X_i'\hat{\beta}))'] \\
&= E[X_i'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X_i] \\
&= X_i' V(\hat{\beta})X_i \\
&= \sigma^2 X_i' (X'X)^{-1} X_i \\
&= \sigma^2 h_{ii};\ i = 1(1)n
\end{aligned}
$$

เราเรียก $h_{ii} = X_i'(X'X)^{-1}X_i$ ว่า **leverage** ของค่าสังเกตชุดที่ i h_{ii} จะบอกให้เราทราบว่าค่าสังเกตชุดที่ i เป็น outlier หรือไม่ ทั้งนี้เพราะ $X_i'(X'X)^{-1}X_i$ เป็น quadratic form เป็น distance จากจุดที่ i ไปยังจุดที่เป็น mean ถ้า h_{ii} มีค่าสูงแสดงว่าค่าสังเกตชุดที่ i อยู่ห่างจาก mean point หรือแตกกลุ่ม เราจะศึกษาเรื่อง leverage อีกครั้ง

แสดงว่า $V(e_i) = \sigma^2 - \sigma^2 h_{ii} = \sigma^2(1 - h_{ii});\ i = 1(1)n$

ดังนั้น $ZPRED_i = \dfrac{\hat{Y}_i - \bar{Y}}{\sqrt{s^2 h_{ii}}}$ โดยที่ $s^2 = MSE = \sigma^2$

อนึ่ง เมื่อพิจารณาเวคเตอร์ $\hat{Y}$ (predicted vector) จะพบว่า

$$
\begin{aligned}
V(\hat{Y}) &= E[(\hat{Y} - E(\hat{Y}))(\hat{Y} - E(\hat{Y}))'] \\
&= E[(X\hat{\beta} - X\beta)(X\hat{\beta} - X\beta)'] \\
&= E[X(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X'] \\
&= XV(\hat{\beta})X' \\
&= \sigma^2 X(X'X)^{-1}X' \\
&= \sigma^2 H
\end{aligned}
$$

ซึ่งเราพิสูจน์ได้ว่า สมาชิกที่ (i, i) ที่เมตริกซ์ $\sigma^2 H$ ก็คือ $V(\hat{Y}_i)$ โดยที่ $V(\hat{Y}_i) = \sigma^2 X_i'(X'X)^{-1}X = \sigma^2 h_{ii}$ เราเรียก $s\sqrt{h_{ii}}$ ว่า standard error of predicted value ตัวที่ i (คือ **SEPRED** ในตารางข้างบน)

อนึ่ง ในการพยากรณ์ (prediction) นั้น เราทำได้ 2 ทาง คือพยากรณ์ค่าเฉลี่ย (คือ E(Y)) กับพยากรณ์ค่า individual Y

ให้ $\hat{Y}_F$ คือค่าพยากรณ์ของ individual Y จะพบว่า

$\hat{Y}_F = X_i\hat{\beta}$ ตรงกันกับ $\hat{Y}_i$ (ซึ่งก็คือ $E(Y_i)$ หรือ $E(Y_i|X_i)$) สิ่งที่แตกต่างคือ variance กรณี individual $\hat{Y}$ จะพบว่า

$$
\begin{aligned}
V(\hat{Y}_F) &= E[(\hat{Y}_F - Y_F)(\hat{Y}_F - Y_F)'] \\
&= E[\{X_i'\hat{\beta} - (X_i'\beta + u_i)\}\{X_i'\hat{\beta} - (X_i'\beta + u_i)\}'] \\
&= E[\{X_i'(\hat{\beta} - \beta) - u_i\}\{X_i'(\hat{\beta} - \beta) - u_i\}'] \\
&= X_i'E(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X_i + E(u_i^2) \\
&= \sigma^2 X_i'(X'X)^{-1}X_i + \sigma^2 \\
&= \sigma^2\{1 + X_i'(X'X)^{-1}X_i\}
\end{aligned}
$$

อย่างไรก็ตาม ในการพยากรณ์และการวิเคราะห์ residual นั้น โดยเจตนาหลักก็คือเราอยากทราบพฤติกรรมของตัวแปรสุ่ม u แต่เพราะเหตุว่า u เป็น unobservable random variable เราจึงศึกษาจาก residual คือ e ซึ่งเป็นตัวแทนของ x แทนกันไป ประเด็นที่จะต้องวิเคราะห์ (เรียกว่า residual analysis) ก็คือ

1. residual ตัวใดมีลักษณะผิดปกติ คือสูงผิดปกติ หรือต่ำผิดปกติ (เรียกว่า extreme หรือ outlying) ตัวใดมีอิทธิพลมากเป็นพิเศษ (เรียกว่า influential) ซึ่งจะโยงกลับไปว่าค่าสังเกตชุดใดผิดปกติ หรือมีอิทธิพลมากเป็นพิเศษ

2. จาก residual ที่มีทั้งสิ้น n ค่า เราสามารถนำไปใช้ศึกษาและวิเคราะห์ว่าข้อตกลงต่าง ๆ เป็นจริงหรือไม่

print out ข้างต้นจะให้ข้อมูลต่าง ๆ เพื่อสนองการศึกษาประเด็นที่ 1 ดังนี้

(1) * PRED ก็คือ $\hat{Y}_i$ (มีเครื่องหมาย * กำกับไว้เพื่อแสดงให้ทราบว่า PRED เป็น tempory variable); $i = 1, 2, \ldots, n$

(2) * ZPRED หมายถึง standardized predictred value

$$
* \text{ZPRED}_i = \frac{\hat{Y}_i - \bar{Y}}{\sqrt{s^2 h_{ii}}} ; i = 1, 2, \ldots, n
$$

(3) * SEPRED หมายถึง standard error of predicted value

$$
* \text{SEPRED}_i = \sqrt{s^2 h_{ii}} ; i = 1, 2, \ldots, n
$$

(4) * ADJPRED หมายถึง adjusted predicted value เรื่องนี้มีคำอธิบายทางทฤษฎีที่พึงทำความเข้าใจเสียก่อนดังนี้

ในการวิเคราะห์สมการถดถอยนั้น ค่าสังเกตบางวาระอาจมีอิทธิพลสูงผิดปกติซึ่งไม่ได้หมายความว่าข้อมูลผิด แต่อาจเพราะมีเหตุการณ์พิเศษบางอย่าง เช่น ความเปลี่ยนแปลงทางการเมือง การเปลี่ยนแปลงทางเศรษฐกิจ สังคม หรือเหตุการณ์ทางธรรมชาติ เช่น น้ำท่วม ฝนแล้ง แผ่นดินถล่ม หรืออาจกลับกัน เช่น น้ำไม่ท่วม ฟ้าฝนบริบูรณ์ แผ่นดินไม่ถล่ม เกิดขึ้นในวาระนั้น ซึ่งจะมีผลให้ค่าสังเกตมีธรรมชาติแปลกไปจากชุดอื่น ๆ หากนำ

ค่าสังเกตชุดนี้ออกจากการวิเคราะห์ จะทำให้ผลการวิเคราะห์ผิดไปจากเดิมอย่างมาก เราเรียกค่าสังเกตชุดนี้ว่า influential observation ในชั้นปฏิบัติแล้วเราไม่อาจทราบได้ว่าค่าสังเกตชุดใดเป็น influential observation เพราะค่าสังเกตนี้อาจแอบหรือไม่แสดงตัว คือเราตรวจดูด้วย outlier แล้วไม่พบ (outlier คือ standaidized residual ที่มีค่าปรากฏนอก 95% CI หรือ 99% CI ของ e หรือในชั้นปฏิบัติก็คือถือว่ามีค่าน้อยกว่า -3.0 หรือมากกว่า $+3.0$) เราจึงถือว่าค่าสังเกตทุกชุดมีสิทธิ์เป็น influential observation ด้วยกันทั้งสิ้น และดังที่กล่าวแล้วข้างต้นว่า ค่าสังเกตที่เป็น influential observation นั้นดูได้จากการทดลองไล่ค่าสังเกตชุดนั้นออกจากการวิเคราะห์ ถ้าเป็น influential observation จริง $\hat{\beta}$ และ $\hat{V}(\hat{\beta})$ จะผิดเพี้ยนไปจากเดิมมาก ค่าพยากรณ์ก็จะผิดปกติมาก influential observation จึงอาจเป็นค่าสังเกตที่มีค่ายิ่ง หรือไร้ค่ายิ่งก็ได้ การจะสงวนไว้หรือคัดทิ้งไปจึงอยู่ที่ความสมเหตุสมผล คำว่าไร้ค่ายิ่งหมายความว่าเราตรวจสอบเหตุการณ์แวดล้อมดีแล้วว่าเป็นเหตุการณ์ปกติเช่นเดียวกับวาระใกล้เคียง ค่าสังเกตผิดปกตินำจะเพราะบันทึกข้อมูลผิด อย่างนี้เรียกว่าไร้ค่ายิ่ง

การคำนวณหา influential observation มีวิธีการดังต่อไปนี้ เรื่องนี้นักศึกษาต้องสนใจให้มาก เพราะพัวพันกับอีกหลายเรื่อง

ขั้นที่ 1 ดึงค่าสังเกตชุดที่ t ออกแล้ววิเคราะห์สมการถดถอยจากค่าสังเกต $(n-1)$ ชุดที่เหลืออยู่ ได้ $\hat{\beta}^{(t)} = [\hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)}, \ldots, \hat{\beta}_k^{(t)}]'$ และ $\hat{Y}_i^{(t)} = \hat{\beta}_1^{(t)} + \hat{\beta}_2^{(t)}X_{2i} + \ldots + \hat{\beta}_k^{(t)}X_{ki}$ หรือ $\hat{Y}_i^{(t)} = X_i'\hat{\beta}^{(t)}$

คือสมการประมาณค่าของ $Y = X\beta + u$ เมื่อค่าสังเกตชุดที่ t ถูกผลักออกจากกลุ่มค่าสังเกต

ขั้นที่ 2 คำนวณหาค่าพยากรณ์ของ Y_i โดยอาศัยสมการ $\hat{Y}_i^{(t)} = X\hat{\beta}^{(t)}$

$$\text{ได้ } \hat{Y}_i^{(t)} = X_i'\hat{\beta}^{(t)} = (1, X_{2i}, X_{si}, \ldots, X_{ki}) \begin{bmatrix} \hat{\beta}_1^{(t)} \\ \hat{\beta}_2^{(t)} \\ \vdots \\ \hat{\beta}_k^{(t)} \end{bmatrix}$$

เรียก $\hat{Y}_i^{(t)}$ ว่า adjusted predicted value หรือ ADJPRED ซึ่งก็คือค่าพยากรณ์ของ Y ณ วาระที่ i ที่คำนวณขึ้นจากสมการถดถอยที่วิเคราะห์จากข้อมูลที่เราได้ตัดค่าสังเกตชุดที่ t ทิ้งไปก่อนที่จะวิเคราะห์

ค่าพยากรณ์ของ Y ณ วาระที่ t คือ $\hat{Y}_t^{(t)} = X_t'\hat{\beta}^{(t)}$ และ $e_t^{(t)} = Y_t - \hat{Y}_t^{(t)}$ เรียก $e_t^{(t)}$ ว่า deleted residual หรือ DRESID ค่าสังเกตชุดใดให้ deleted residual สูง ทำให้พอจะเดาได้ว่าค่าสังเกตชุดนั้นอาจเป็น influential observation

เมื่อศึกษาถึงตรงนี้เราจึงพบว่า เราน่าจะมีตัวสถิติที่ช่วยบอกอะไรให้ทราบได้รวม ๆ อีก 1 ตัว คือ Cook's distance

ขั้นที่ 3 เมื่อ $i = 1, 2, \ldots, n$ และ $t = 1, 2, \ldots, n$ เราย่อมคำนวณหา ADJPRED ได้ครบ และจากข้อมูล n ชุด เราวิเคราะห์สมการถดถอยตามปกติ ได้ $Y = X\hat{\beta}$, s^2 และ $\hat{Y}_i = X_i'\hat{\beta}$ และคำนวณหา Cook's distance ได้ดังนี้

$$C_t = \sum_{i=1}^{n} (\hat{Y}_i^{(t)} - \hat{Y}_i)^2/ks^2; t = I. 2. \ldots, n$$

ค่าสังเกตชุดใดให้ Cook's distance สูง แสดงว่าค่าสังเกตชุดนั้นเป็น influential observation

คราวนี้เรากลับมาดู residual statistics ตามตารางที่ผ่านมากันต่อไป ซึ่งเป็นเรื่องไม่ยาก เพราะทราบเหตุผลและการพัฒนาทางทฤษฎีมาแล้วดังนี้

(5) * RESID หมายถึง residual e กล่าวคือ $e_i = Y_i - \hat{Y}_i$; $i = 1, 2, \ldots, n$

(6) * ZRESID หมายถึง standardized residual

$$ZRESID_i = e_i/\sqrt{MSE}; s^2 = MSE \text{ จากสมการ } Y = X\hat{\beta}$$

ZRESID เกิดจากการนำเอา e_i หารด้วย SE (คือ $\sqrt{MSE}$ หรือ s) อาจมีจุดอ่อนอยู่เล็กน้อยตรงที่ s^2 เป็นค่า variance ร่วมกันของ $e_1, e_2, \ldots, e_n$ เมื่อนำ s มาถ่วงน้ำหนักให้แก่ $e_1, e_2, \ldots, e_n$ ซึ่งมีค่าที่แตกต่างกัน ซึ่งน่าจะทำให้มองไม่เห็นเหตุการณ์จำเพาะวาระได้ จึงมีการสร้าง residual ตัวใหม่ขึ้นเรียกว่า studentized residual (*SRESID)

$$(7)\ SRESID_i = e_i/\sqrt{\hat{V}(e_i)}; i = 1, 2, \ldots, n$$
$$= e_i/\sqrt{s^2(1 - h_{ii})}$$
$$= e_i/\sqrt{s^2(1 - X_i'(X'X)^{-1}X_i)}$$

โดยปกติ ZRESID กับ SRESID จะมีค่าใกล้เคียงกัน แต่ก็ไม่แน่ว่าจะมีเท่ากันหรือใกล้เคียงกันเสมอไป ข้อแตกต่างก็มีเพียงเฉพาะ SRESID ช่วยให้ติดตามจับความผันแปรของ u_i ที่แปรไปตามวาระได้ดีกว่า ZRESID

(8) * DRESID หมายถึง deleted residual ค่านี้คำนวณจาก ADJPRED ดังนี้

$$DRESID_t = e_t^{(t)} = Y_t - \hat{Y}_t^{(t)}; t = 1, 2, \ldots, n$$

(9) * SDRESID หมายถึง studentized deleted residual คำนวณต่อเนื่องจาก DRESID ดังนี้

$$SDRESID_t = e_t^{(t)}/s(e_t^{(t)})$$

โดยที่ $s^2(e_t^{(t)}) = MSE_{(t)}[1 + X_t'\{X_{(t)}'X_{(t)}\}^{-1}X_t]$

และเราสามารถพิสูจน์ได้ว่า

$$d_i^* = SDRESID_i = e_i \left[\frac{(n-1)-k}{SSE(1-h_{ii})-e_i^2} \right]^{1/2}; \; i = 1, 2, \ldots, n$$

$$d_i^* \sim t_{(n-k-1)}$$

โดยที่ SSE, e_i และ h_{ii} คำนวณจากกรณีปกติคือกรณีมีค่าสังเกตครบ n ค่าหมายความว่าเราไม่จำเป็นต้องทำ loop โดยสลับการผลักค่าสังเกตต่าง ๆ ออกจากกลุ่มค่าสังเกตเพื่อหา $\hat{Y}_i^{(i)}$ และ $e_i^{(i)}$ แต่เราคำนวณได้โดยตรงจากการวิเคราะห์ตามปกติเมื่อมีค่าสังเกตครบ n ชุด

(10) * MAHAL หมายถึง Mahalanobis distance ใช้สำหรับแสดงให้เห็นว่าค่าสังเกตใดมากหรือน้อยผิดปกติหรืออยู่ห่างจาก mean point คือ $(\bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k)$

$$MAHAL_i = \sum_{j=2}^{k} \left(\frac{X_{ji} - \bar{X}_j}{s_j} \right)^2; \; i = 1, 2, \ldots, n$$

ค่าสังเกตชุดใดให้ค่า Mahalanobis distance สูง แสดงว่าค่าสังเกตชุดนั้นมีตัวแปร X ที่ให้ค่าผิดปกติปนอยู่ จำเป็นต้องตรวจสอบ แต่จะคัดทิ้งหรือไม่ให้ดูที่เหตุผล บ่อยครั้งที่เราไม่ได้ Mahalanobis distance ที่สูงต่ำผิดปกติเลย ซึ่งก็มิได้แปลว่าเราจะไม่มีปัญหา จำเป็นต้องตรวจต่อไปว่ามีค่าสังเกตใดที่มีพลังสูงผิดปกติ (คือเป็น influential observation) ปนอยู่หรือไม่โดยการตรวจดูด้วย Cook's distance กรณีนี้ Mahalanobis distance ตรวจไม่ได้ (สรุปว่า HAHAL ใช้ดูว่าค่าสังเกตชุดใดผิดปกติ โดยดูที่ X's ขณะที่ COOK D ใช้ดูว่าค่าสังเกตใดผิดปกติ โดยดูที่ Y)

(11) * COOK D หมายถึง Cook's distance เป็นตัวสถิติที่ใช้ตรวจดูว่าค่าสังเกตชุดใดเป็น influential case

$$C_t = \sum_{i=1}^{n} (\hat{Y}_i^{(t)} - \hat{Y}_i)^2 / ks^2; \; t = 1, 2, \ldots, n$$

โดยที่ s^2 คือ MSE จากสมการ $Y = X\hat{\beta}$, k = จำนวนพารามิเตอร์ในสมการ และเราสามารถพิสูจน์ได้ว่า

$$c_i = \frac{e_i^2}{ks^2} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]; \; i = 1, 2, \ldots, n$$

โดยที่ e_i, h_{ii} และ s^2 คำนวณจากการวิเคราะห์ตามปกติเมื่อมีค่าสังเกตครบ n ชุด

(12) * LEVER หมายถึง leverage h_{ii} เราใช้ h_{ii} เป็นดัชนีช่วยชี้ว่าค่าสังเกตชุดใดสูงต่ำผิดปกติ คือ เบนห่างจาก mean point คือจุด $(\bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k)$ มากน้อยเพียงใด (มองกันในเทอมของตัวแปร X เหมือน MAHAL) ค่าสังเกตใดให้ค่า h_{ii} สูงแสดงว่าค่าสังเกตนั้นเป็น extreme case การจะคัดค่าสังเกตชุดนี้ทิ้งหรือไม่จึงต้องดูที่เหตุผล

Leverage ใช้ช่วยการวิเคราะห์ residual ช่วยอย่างไร? เรื่องนี้ต้องดูเหตุผลทางทฤษฎี

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

แสดงว่า $\hat{Y}_i = h_{ii}Y_i;\ i = 1, 2, \ldots, n$ .....(1)

สมการ (1) ชี้ให้เห็นว่า ถ้า h_{ii} มีค่ามาก แสดงว่า Y_i มีผลต่อ $\hat{Y}_i$ มาก แต่ $h_{ii} = X_i'(X'X)^{-1}X_i$ จึงแปลว่า X_i มีผลต่อ $\hat{Y}_i$ มาก ถ้า h_{ii} มีค่ามาก

แต่ $V(e_i) = \sigma^2(1 - h_{ii})$ หรือ $\hat{V}(e_i) = s^2(1 - h_{ii})$ .....(2)

จากสมการ (2) และเราทราบว่า

(1) $0 \leq h_{ii} \leq 1$

(2) $\sum_{i=1}^{n} h_{ii} = 1$

(3) $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}$

ถ้า $h_{ii} \to 1$ จะทำให้ $V(e_i)$ มีค่าต่ำ หรือ $\hat{Y}_i \to Y_i$ เรื่องนี้ชี้ให้เห็นว่า ค่าสังเกตที่ผิดปกติอาจตรวจไม่พบด้วย residual analysis ตามวิธี outlier เพราะค่าพยากรณ์ของ Y_i วางบนเส้นสมการประมาณค่าหรือใกล้เคียงจนมองไม่เห็นความผิดปกติ

การใช้ leverage นั้นให้ดูว่า $h_{ii} > 2\bar{h}$ หรือไม่ ถ้า $h_{ii} > 2\bar{h}$ ให้ถือว่าค่าสังเกตชุดที่ i ผิดปกติ (extreme)

รายละเอียดเรื่องเหล่านี้สามารถวิเคราะห์เพิ่มเติมได้จากการดูผลจาก / CASEWISE ผลจากการสั่ง / RESIDUALS = HISTOGRAM OUTLIERS (*DRESID) DURBIN ทำให้ได้ปรากฏผลลัพธ์ดังนี้

keyword DURBIN จะแสดงค่าของ durbin-watson statistics คือ

$$\text{durbin-watson} = \sum_{t=2}^{n}(e_t - e_{t-1})^2 / \sum_{t=1}^{n} e_t^2$$

สถิติตัวนี้ใช้ตรวจสอบดูว่าข้อตกลงว่า $E(u_iu_j) = 0$ เป็นจริงหรือไม่ ถ้าไม่เป็นจริงแสดงว่าเกิดปัญหา autocorrelation

keyword OUTLIERS (temp vars) จะแสดงค่า extreme ของตัวแปรชั่วคราว ซึ่งเรากำหนดเองตามต้องการให้เห็น 10 ค่า ในที่นี้เราระบุให้ตัวแปรชั่วคราวเป็น DRESID โปรแกรมจึงแสดง deleted residual รวม 10 ค่าเรียงตามขนาด (ไม่นับเครื่องหมาย) จากมากไปหาน้อย

Page 12 SPSS/PC+

Outliers – Deleted (Press) Residual

Case #	* DRESID
8	– 6.32637
6	4.26840
1	3.35531
5	3.24177
4	– 3.21002
10	– 3.12068
3	– 2.69474
2	2.32625
12	2.11918
7	– 1.60882

keyword HISTOGRAM ให้ผลเป็นกราฟแท่งดังภาพ บังเอิญเรามีข้อมูลน้อยเกินไป กราฟจึงไม่เป็นแท่งที่เด่นชัด เรื่องนี้มีคำอธิบายประกอบดังนี้

Page 13 SPSS/PC+

Histogram – Standardized Residual

N	Exp N	(* = 1 Cases,	. : = Normal Curve)
0	.01	Out	
0	.02	3.00	
0	.05	2.67	
0	.12	2.33	
0	.24	2.00	
0	.43	1.67	
1	.71	1.33	:
1	1.05	1.00	:
3	1.38	.67	:**
1	1.63	.33	*.
1	1.72	0.0	*.
2	1.63	–.33	*:
1	1.38	–.67	:
2	1.05	– 1.00	:*
0	.71	– 1.33	.
0	.43	– 1.67	
1	.24	– 2.00	*
0	.12	– 2.33	
0	.05	– 2.67	
0	.02	– 3.00	
0	.01	Out	

* หมายถึงค่าสังเกต (observed frequency)

. หมายถึงค่าคาดหมายที่คำนวณมาจาก Normal Distribution

: หมายถึงกรณีที่ค่าคาดหมายกับค่าสังเกตเท่ากันหรือใกล้เคียงกัน

N หมายถึง observed frequency

Exp N หมายถึง expected frequency

Out หมายถึงค่าของ residual (โดยปกติจะใช้ standardized residual) ที่มีค่าปรากฏอยู่ภายนอก 95% CI คือน้อยกว่า − 3.00 และมากกว่า + 3.00

เป้าหมายของของเรื่องนี้ก็คือเพื่อจะดูว่า u มีการแจกแจงแบบปกติหรือไม่ ซึ่งเราดูได้จากการทำตารางแจกแจงความถี่ของ standardized residual แล้วนำเอาความถี่คาดหมาย (expected frequency) กับความถี่จริง (observed frequency) ของจำนวน standardized residual ที่จะพึงปรากฏอยู่ในแต่ละช่วงซึ่งกว้างประมาณ 0.333 มาเปรียบเทียบกัน (เนื่องจาก standardized residual ประมาณ 99 ส่วน ใน 100 ส่วน จะมีค่าปรากฏในช่วงระหว่าง − 3.00 ถึง + 3.00) โดยให้ซอยเป็นช่วงย่อยกว้างช่วงละประมาณ 0.333 ได้ 18 ช่วง รวมจุดช่วงกลางอีกช่วงหนึ่ง รวมเป็น 19 ช่วง คือ (− 3.00 − 2.67), (− 2.66 − 2.34), (− 2.33 − 2.00), ..., (− 0.33 − 0.00), (0.00 − 0.33), ..., (2.67 − 3.00) เพิ่มอีก 2 ช่วงหัวและท้าย คือ out รวมเป็น 21 ช่วง แต่ละช่วงเหล่านี้ให้คำนวณหาค่าความน่าจะเป็นจาก normal distribution

$$p_i = \Pr\{a < Z < b\}$$

เช่น $p_1 = \Pr\{Z < -3.00\}$

$p_2 = \Pr\{-3.00 < Z < -2.67\}$

เป็นต้น

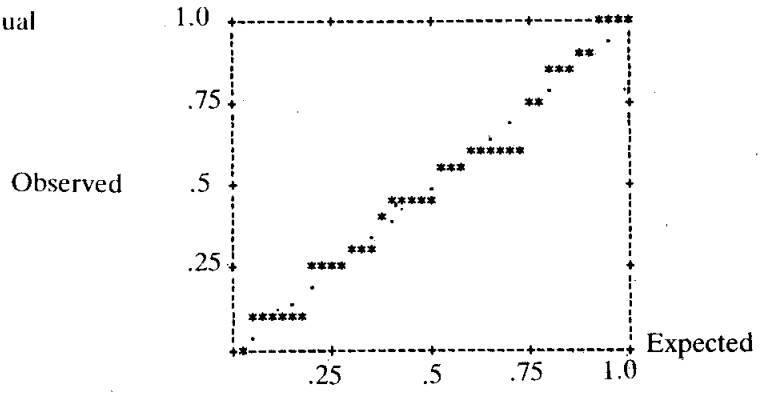
$E_i = np_i$; $i = 1, 2, \ldots, 21$ คือความถี่คาดหมาย (Exp N)

O_i = จำนวนความถี่จากการนับรอยคะแนน (tally) (คือ N ในตาราง)

เมื่อนำเอา N และ Exp N มาพล็อตกราฟแท่งก็จะทราบได้ว่าความถี่จริงมีธรรมชาติสอดคล้องกับความถี่ที่พึงมีได้ตาม normal distribution หรือไม่ ถ้า $O_i = E_i$ (หรือ Exp N กับ N เท่ากัน) แสดงว่าความถี่จริงตรงกันกับความถี่ที่พึงได้จาก normal distribution

อีกวิธีหนึ่งก็คือ นำเอา observed probability มาพล็อตร่วมกัน expected probabity ถ้าค่าความน่าจะเป็นเท่ากันซึ่งดูได้จากเส้นตรง 45% แสดงว่า u มีธรรมชาติสอดคล้องกับ normal distribution (ผลจาก keyword NORMPROB) ดังภาพ

Normal Probability (P–P) Plot
Standardized Residual



คำสั่ง / CASEWISE จะทำการพล็อต standardized residual ให้เราตรวจดูว่าค่าสังเกตชุดใดเป็น outlier (คือตกอยู่นอกช่วงเชื่อมั่น 99%) keyword ALL หมายถึงสั่งให้พล็อต residual ทุกค่า และถ้าเราประสงค์จะให้แสดงค่าตัวแปรอื่นก็ให้ระบุลงไปด้วย (ดูรายละเอียดคำสั่งนี้ท้ายเล่ม) ในที่นี้สั่งให้แสดงค่าตัวแปรตาม (X5 = Y) ค่าพยากรณ์ ($\hat{Y}$) และ residual (e) ผลการพล็อตไม่พบ outlier

* * * * MULTIPLE REGRESSION * * * *

Equation Number 1 Dependent Variable... X5
Casewise Polt of Standardized Residual
* : Selected M : Missing

Case #	− 3.0 0.0 3.0	X5	* PRED	* RESID
1	0 : . . . . . : . . . . . : 0	79	76.3399	2.1601
2	. . * .	74	72.6118	1.6882
3	. . * .	104	106.6579	− 2.3579
4	. * . .	88	90.0811	− 2.4811
5	. * . .	96	92.9166	2.9834
6	. . * .	109	105.4299	3.7701
7	. . * .	103	103.7335	− 1.0335
.8	. * . .	73	77.5234	− 5.0234
9	. . * .	93	92.4703	.6297
10	. * . .	116	117.3737	− 1.4737
11	. * . .	84	83.6629	.1371
12	. . * .	113	111.5695	1.7305
13	. * . .	109	110.1295	−.7295
Case #	0 : . . . . . : . . . . . : 0	X5	* PRED	* RESID
	− 3.0 0.0 3.0			

การสั่งวิ่งโปรแกรมเพื่อวิเคราะห์ด้วยวิธีอื่น ๆ คือ ENTER BACKWARD STEPWISE REMOVE และ TEST จะให้ผลลัพธ์ที่อาจแตกต่างจากที่กล่าวมา แต่การแปลผลก็เป็นไปในทำนองเดียวกัน

แบบฝึกหัดบทที่ 4

1. จากการศึกษาการดำเนินงานของสถานประกอบการ 89 แห่ง พบว่ามีข้อมูลเกี่ยวกับค่าใช้จ่าย (X_1) อัตราการผลิต (X_2) และอัตราการขาดงานของพนักงาน (X_3) ดังนี้ (อะเรย์ที่กำหนดให้คือ X´X)

$\bar{X}1 = 5.8, \quad \bar{X}2 = 2.9, \quad \bar{X}3 = 3.9$

	X1	X2	X3
X1	113.6	36.8	39.1
X2		50.5	-66.2
X3			967.1

จงวิเคราะห์ความสัมพันธ์ระหว่าง X_1 กับ X_2 และ X_3 สร้างตาราง ANOVA เพื่อศึกษาอิทธิพลของตัวแปรแต่ละตัว (เฉพาะ X_2 หรือ X_3) และทั้งสองตัว

2. ให้ X1 = ค่า log ของปริมาณการบริโภคสินค้าประเภทอาหารต่อบุคคล
   X2 = ค่า log ของราคาสินค้าประเภทอาหาร
   X3 = ค่า log ของรายได้ต่อบุคคล

กำหนดให้อะเรย์ VC-matrix ปรากฏดังนี้

	X1	X2	X3
X1	7.59	3.12	26.99
X2		29.16	30.80
X3			133.00

และสมมุติว่า สมการความสัมพันธ์คือ $Y_1 = AY^{\alpha}_2 Y^{\beta}_3$ (ในที่นี้ $X_1 = \log Y_1$ $X_2 = \log Y_2$, $X_3 = \log Y_3$) จงประมาณค่าสมการถดถอยพร้อมช่วงเชื่อมั่น 95% ของ α และ β กำหนดให้ n = 20

3. จากการศึกษาเรื่องค่าใช้จ่ายในการทำเหมืองลิกไนต์ ข้อมูลของตัวแปรค่าใช้จ่ายต่อตัน (X_1) ดีกรีของการใช้เครื่องจักรหรือความทันสมัย (X_2) ค่าต่อเลขทางธรณีวิทยาแสดงความยาก

ในการขุด (X_3) และร้อยละของพนักงานที่ขาดงาน (X_4) และให้

$$Z_i = \log X_i \; ; \; i = 1, 2, 3, \quad Z_4 = X_4$$

และให้อะเรย์แสดงสหสัมพันธ์ระหว่างตัวแปรปรากฏดังนี้

	Z1	Z2	Z3	Z4
Z1	1.00000	0.3597	0.5749	0.4109
Z2		1.0000	0.4630	0.3050
Z3			1.0000	0.2702
Z4				1.0000

จงวิเคราะห์สมการถดถอย กำหนดให้ n = 62

4. ในบางกรณีเราจะต้องมีข้อจำกัดบางประการเกี่ยวกับ สปส.ความถดถอย เช่นในกรณีการผลิต $Y = AL^{\alpha}K^{\beta}$ (Y = ผลผลิต L = จำนวนแรงงาน K = จำนวนเงินทุน) ซึ่งจะต้องมีเงื่อนไขว่า $\alpha+\beta = 1$ หรือ $[1 \quad 1]\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix} = 1$ หรือ $R\hat{\beta} = r$ เป็นต้น ดังนั้น สมการ $R\hat{\beta} = r$ จึงเป็น constraint condition ของฟังก์ชันเป้าหมายคือ $\Sigma e^2 = (Y-X\hat{\beta})'(Y-X\hat{\beta})$ ดังนั้น Lagrange function คือ

$$F = (Y-X\hat{\beta})'(Y-X\hat{\beta})-\theta'(R\hat{\beta}-r)$$

จงประมาณค่า β (คือหาค่า $\hat{\beta}$)

5. จากสมการ $Y = \beta_1+\beta_2X_2+\beta_3X_3+u$ โดยมีข้อจำกัดว่า $\beta_2 = \beta_3$ และจากข้อมูล 23 รายการพบว่า

$$X'X = \begin{bmatrix}20 & 0\\ 0 & 40\end{bmatrix} \qquad X'Y = \begin{bmatrix}15\\ 25\end{bmatrix}$$

6. coviance matrix ระหว่างตัวแปร x_i ; i = 0, 1, 2, 3 ปรากฏดังนี้

	x0	x1	x2	x3
x0	13	12	15	27
x1	12	14	10	24
x2	15	10	16	26
x3	27	24	26	50

จงวิเคราะห์สมการถดถอย $x_0 = f(x_1, x_2, x_3)$ แล้วพยากรณ์ค่า x_0 สำหรับอัตราเปลี่ยนดังนี้

ก. $\Delta x_1 = 1, \Delta x_2 = -1, \Delta x_3 = 0$

ข. $\Delta x_1 = 1, \Delta x_2 = -1, \Delta x_3 = 2$

7. ข้อมูลต่อไปนี้คือข้อมูลแสดงยอดขายอัญมณีของปีต่าง ๆ รวม 16 ไตรมาส ดังนี้

	T ไตรมาส/ปี	S (1,000,000)	Q4	Q3	Q2
1990	1	24	0	0	0
	2	29	0	0	1
	3	29	0	1	0
	4	50	1	0	0
1991	5	24	0	0	0
	6	30	0	0	1
	7	29	0	1	0
	8	51	1	0	0
1992	9	26	0	0	0
	10	29	0	0	1
	11	30	0	1	0
	12	52	1	0	0
1993	13	25	0	0	0
	14	30	0	0	1
	15	29	0	1	0
	16	50	1	0	0

1. จงวิเคราะห์สมการถดถอย $S = \beta_1 + \beta_2 T + u$
2. สังเกตจากข้อมูลจะเห็นว่าข้อมูลมีค่าสูงผิดปกติทุกไตรมาสที่ 4 จงแก้ปัญหาฤดูกาล (deseasonalization) 3 วิธีดังนี้

1) ผนวก Q4 เข้าไปในข้อ 1.

2) ผนวก Q2,Q3,Q4 เข้าไปในข้อ 1.

3) วิเคราะห์สมการถดถอย $S = \beta 1+\beta_2 T+ u$ โดยก่อนวิเคราะห์ให้แปลงข้อมูลของตัวแปร S โดยวิธีค่าเฉลี่ยเคลื่อนที่กำหนดให้ระยะเฉลี่ย (average length) เท่ากับ 4

4) วิเคราะห์สมการถดถอยในข้อ 3) โดยแปลงข้อมูลของตัวแปร S โดยวิธีค่าเฉลี่ยเคลื่อนที่ 2 ครั้ง (double moving average)

8. ผลการทดสอบความสิ้นเปลืองน้ำมันเชื้อเพลิงของรถยนต์ 6 คันปรากฏดังนี้

	ความสิ้นเปลือง (ไมล์/แกลลอน) M	จำนวนแรงม้าของเครื่องยนต์ HP
รถยี่ห้อ A	21	210
	18	240
	15	310
รถยี่ห้อ B	20	220
	18	260
	15	320

จงวิเคราะห์สมการถดถอย M = f (HP) โดยเปรียบเทียบระหว่างยี่ห้อ พร้อมวาดภาพแสดงกราฟของสมการยี่ห้อ A และยี่ห้อ B

(แนะนำ ให้ $X = \begin{cases} 1 & \text{ถ้ายี่ห้อ A} \\ 0 & \text{ถ้ายี่ห้อ B} \end{cases}$

ดังนั้น $M = \beta_1 + \beta_2 HP + \beta_3 X + u$

9. จำนวนปีที่อยู่ในสถานศึกษาของนักเรียนและรายได้ของบิดาจำแนกตามถิ่นที่อยู่ (เมือง, ชนบท) ปรากฏดังนี้

ถิ่นที่อยู่	จำนวนปี ที่อยู่ในโรงเรียน E	รายได้ของ บิดา F
เขตเมือง	15	8,000
	18	11,000
	12	9,000
	16	12,000
เขตชนบท	13	5,000
	10	3,000
	11	6,000
	14	10,000

จงวิเคราะห์สมการถดถอย

10. จำนวนประชากรใน t ปี (t = 0 หมายถึงปีเริ่มต้น) ปรากฏดังนี้

เวลา t (ปี)	จำนวนประชากร
0	1,570
20	1,810
40	2,250
50	2,510

1. จงวาดกราฟ
2. สมการใดต่อไปนี้เหมาะสมที่สุด ให้พิจารณาความง่ายในการนำไปใช้ ความแม่นยำในการพยากรณ์ ความสมเหตุสมผลของการพยากรณ์

1) $Y = \alpha+\beta t$

2) $Y = \alpha+\beta t+\gamma t^2$

3) $Y = \alpha e^{\beta t}$

<u>หมายเหตุ</u> ถ้า Y = ค่าใช้จ่ายในการผลิต Q = ปริมาณการผลิต
สมการ $Y = \alpha+\beta Q+rQ^2$ เรียกว่า Cost function จะมีรูปของโค้งเป็นตัว U

11. จงอธิบายวิธีดำเนินการเพื่อประมาณค่าสัมประสิทธิ์ความถดถอยของสมการต่อไปนี้

1) $Y = \alpha+\beta X$

2) $Y = \alpha+\beta X+\gamma e^{st}$

3) $Y = \alpha+\beta t+\gamma \text{Sin}\ (2\pi t/12)$

<u>หมายเหตุ</u> βt แสดง linear growth, sin $(2\pi t/12)$ แสดง annual cycle)

4) $Y = \alpha e^{\beta t}$

5) $Y = \alpha(1+\beta)^t$

<u>หมายเหตุ</u> β หมายถึง growth rate

6) $Y = \alpha\beta^t\gamma^X$

7) $Y = \alpha\beta^t+\gamma^X$

12. ให้ β และ β_* เป็นสัมประสิทธิ์ความถดถอยของสมการ $Y = \alpha+\beta X$ และ $X = \alpha_*+\beta_* Y$ จงแสดงให้เห็นว่าข้อต่อไปนี้จริงหรือไม่

1) $\hat{\beta} = r\,(s_y/s_x)$

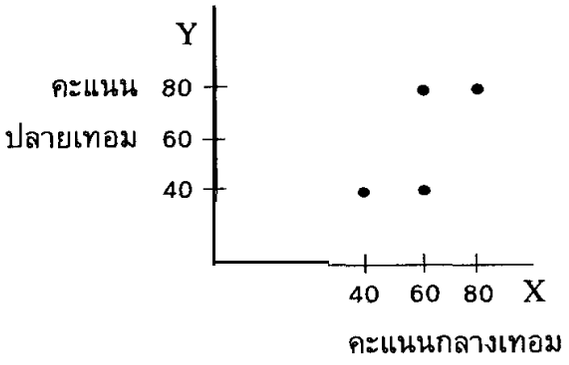
2) $\hat{\beta}_* = r\,(s_x/s_y)$

3) $\hat{\beta}\,\hat{\beta}_* = r^2$

4) ถ้า $\hat{\beta} > 1$ แล้วจำเป็นที่ $\hat{\beta}_* < 1$

5) ถ้า $\hat{\beta} < 1$ แล้วจำเป็นที่ $\hat{\beta}_* > 1$

13. จากภาพต่อไปนี้ซึ่งแสดงคะแนนสอบปลายเทอมและคะแนนกลางเทอม



จงใช้กราฟ (ไม่ต้องคำนวณด้วยสูตร) เพื่อหา

1) สมการถดถอย $Y = \alpha+\beta X$
2) สมการถดถอย $X = \alpha_*+\beta_* Y$
3) ค่าสหสัมพันธ์
4) คำนวณค่าของ Y เมื่อ X = 70
5) คำนวณค่าของ X เมื่อ Y = 70

14. ข้อมูลเกี่ยวกับอาหารปรากฏดังนี้

ปีที่ (t)	Q_D	P_D	Y	Qs	Ps
1	98.6	100.2	87.4	108.5	99.1
2	101.2	101.6	97.6	110.1	99.1
3	102.4	100.5	96.7	110.4	98.9
4	100.9	106.0	98.2	104.3	110.8
5	102.3	108.7	99.8	107.2	108.2
6	101.5	106.7	100.5	105.8	105.6
7	101.6	106.7	103.2	107.8	109.8
8	101.6	108.2	107.8	103.4	108.7
9	99.8	105.5	96.6	102.7	100.6
10	100.3	95.6	88.9	104.1	81.0
11	97.6	88.6	75.1	99.2	68.6
12	97.2	91.0	76.9	99.7	70.9
13	97.3	97.9	84.6	102.0	81.4
14	96.0	102.3	90.6	94.3	102.3
15	99.2	102.2	103.1	97.7	105.0
16	100.3	102.5	105.1	101.1	110.5
17	100.3	97.0	96.4	102.3	92.5
18	104.1	95.8	104.4	104.4	89.3
19	105.3	96.4	110.7	108.5	93.0
20	107.6	100.3	127.1	111.3	106.6

Q_D = ปริมาณการบริโภคสินค้าประเภทอาหาร (ต่อบุคคล)

Q_S = ปริมาณผลผลิตสินค้าประเภทอาหาร (ต่อบุคคล)

P_D = ราคาจำหน่ายปลีกอาหาร/ดัชนีค่าครองชีพ

Y = รายได้/ดัชนีค่าครองชีพ

P_S = ราคารับซื้ออาหารจากเกษตรกร/ดัชนีค่าครองชีพ

(เครื่องหมาย / ในที่นี้หมายถึงการหาร)

จงวิเคราะห์สมการถดถอยพร้อมอภิปรายผล

1) $Q_D = \alpha+\beta P_D$

2) $Q_D = \alpha+\beta Y$

3) $Q_D = \alpha+\beta_1 P_D+\beta_2 Y$

4) $Q_D = \alpha+\beta_1 P_D+\beta_2 Y+\beta_3 t$

5) $Q_S = \alpha+\beta P_S$

6) $Q_S = \alpha+\beta[P_S+(P_S)-1]/2$

7) $Q_S = \alpha+\beta_1(P_S)-1+\beta_2(Q_S)-1$

15. ข้อมูลเกี่ยวกับน้ำมันเชื้อเพลิงปรากฏดังนี้

ปีที	K.	C	M	Pg	Pt	POP	L	Y
1	9732	30.87	14.95	97.0	.533	145	59.3	1513
2	9573	33.39	14.96	100.8	.564	147	60.6	1567
3	9395	36.35	14.92	105.4	.633	150	61.3	1547
4	9015	40.33	14.40	104.3	.678	152	62.2	1646
5	9187	42.68	14.50	98.0	.694	155	62.0	1657
6	9361	43.82	14.27	97.3	.723	158	62.1	1678
7	9370	46.46	14.39	100.0	.765	160	63.0	1726
8	9308	48.41	14.57	101.4	.814	163	63.6	1714
9	9359	52.09	14.53	101.0	.840	166	65.0	1795
10	9348	54.25	14.36	103.3	.860	170	66.6	1839
11	9391	56.38	14.40	103.2	.862	172	66.9	1844
12	9494	57.39	14.30	98.5	.879	175	67.6	1831
13	9529	60.13	14.30	98.1	.897	178	68.4	1881
14	9446	62.26	14.28	98.6	.913	181	69.6	1883
15	9456	63.87	14.38	96.4	.944	184	70.5	1909
16	9441	66.64	14.37	95.0	.965	187	70.6	1968
17	9240	69.84	14.26	93.2	.965	189	71.8	2013
18	9286	72.97	14.25	91.8	.970	192	73.1	2123
19	9286	76.63	14.15	92.6	.972	194	74.5	2235
20	9384	80.11	14.10	92.7	.979	196	75.8	2331
21	9399	82.37	14.05	93.1	1.000	199	77.3	2398
22	9488	85.79	13.91	90.6	1.004	201	78.7	2480
23	9633	89.16	13.75	89.0	1.026	203	80.7	2517

โดยที่ K = ปริมาณแสดงการใช้รถ (ไมล์) ต่อคันต่อปี

C = จำนวนรถยนต์ (1,000,000 คัน)

M = ความสิ้นเปลืองเชื้อเพลิง (ไมล์/แกลลอน)

Pg = ดัชนีราคาจำหน่ายปลีกน้ำมันเชื้อเพลิง/ดัชนีราคาผู้บริโภค

Pt = ดัชนีราคาค่าขนส่ง

POP = จำนวนประชากร (1,000,000 คน)

L = จำนวนกำลังแรงงาน (1,000,000 คน)

Y = รายได้ต่อบุคคล

จงวิเคราะห์สมการถดถอยแล้วอภิปรายผล

16. สมการประมาณค่าต่อไปนี้ใช้ศึกษาว่าภาษีทรัพย์สิน (เช่นภาษีโรงเรือน ภาษีที่ดิน) และค่าใช้จ่ายสาธารณะอื่น ๆ ว่ามีผลต่อค่าของทรัพย์สินนั้นหรือไม่ กำหนดให้ตัวแปรตามคือค่ามัธยฐานของราคาบ้านพักอาศัย

ตัวแปรอิสระคือ

T = อัตราภาษี

E = ค่าใช้จ่ายสาธารณะ (ของรัฐ) เพื่ออุดหนุนการศึกษาถั่วเฉลี่ยต่อนักเรียน 1 คน

Z = ค่าใช้จ่ายของรัฐถั่วเฉลี่ยต่อพลเมือง 1 คน สำหรับด้านอื่น ๆ เช่น การรักษาความสงบ ดับเพลิง สร้างถนน สาธารณสุข ห้องสมุด สวนสาธารณะ ฯลฯ

M= ระยะทางจากเขตย่านการค้าที่ใกล้ที่สุด

R = ค่ามัธยฐานของจำนวนห้องของบ้านพักอาศัย

N = ร้อยละของบ้านพักอาศัยที่ปลูกใหม่

Y = ค่ามัธยฐานของรายได้ครอบครัว

P = ร้อยละของประชากรที่มีรายได้ครัวเรือนต่ำกว่า 3,000$

B = ร้อยละของที่อยู่อาศัยที่มีห้องน้ำเกินกว่า 1 ห้อง

สมการที่ได้ปรากฏดังนี้

สมการที่	C	log T	log E	log M	R	N	Y	P	B	log Z	R^2
1	-23	-9.1	1.4	0.73	3.8	.036	1.8	1.4			.97
	(14)	(3.4)	(2.1)	(0.50)	(1.4)	(.013)	(.50)	(.11)			
2	-49	-8.0	4.3		8.0	.058					.93
	(18)	(4.5)	(2.8)		(.70)	(.016)					
3	-59	-3.8	8.4		4.8	.022			.096		.95
	(16)	(4.3)	(3.0)		(1.6)	(.021)			(.042)		
4	-21	-3.6	3.3	-1.4	1.7	,052	1.5	.30			.93
	(8.5)	(.88)	(1.5)	(.29)	(.42)	(.013)	(.17)	(.083)			
5	-53	-4.2	8.8	-1.5	3.9	.083					.81
	(12)	(1.4)	(2.3)	(.47)	(.50)	(.018)					
6	26	-4.2	3.6	-1.2	1.4	.06	1.5	.30		1.5	.94
	(8.7)	(.91)	(1.5)	(.28)	(.44)	(.014)	(.17)	(.097)		(.79)	

สมการที่ 1-3 คำนวณจากข้อมูลของเมืองชายฝั่ง 19 เมือง ท่านสามารถสรุปอย่างอื่นเพิ่มจากที่สรุปไว้ว่าคนรวยจะอยู่ในบ้านราคาแพงได้หรือไม่ (เพราะมีห้องนอนและห้องน้ำหลายห้อง) สมการ 4-6 คำนวณจากชุมชนต่าง ๆ ในเขตมหานครนิวยอร์คจำนวน 53 ชุมชน จงสรุปผล