ข้อมูลที่เปลี่ยนแปลงค่าไปได้ง่าย จะต้องมีน้ำหนักหรือความน่าเชื่อถือน้อยกว่าข้อมูลที่ค่อนข้างคงที่

นอกจากนี้ความน่าเชื่อถือของข้อมูลยังเกี่ยวข้องอยู่กับหน่วยที่ใช้วัดอีกด้วย

2. ลักษณะความน่าเชื่อถือ

ข้อมูลแต่ละชนิดย่อมมีลักษณะ และขนาดของความน่าเชื่อถือแตกต่างกันไปตามธรรมชาติ ข้อมูลที่เกี่ยวกับเรื่องส่วนตัวที่น่าอับอาย ข้อมูลที่เกี่ยวกับทรัพย์สินหรือผลประโยชน์ และข้อมูลที่เกี่ยวกับความจำ และการคำนวณจะมีความน่าเชื่อถือต่ำ ส่วนข้อมูลที่เกี่ยวข้องกับความผูกพันธ์ ความใกล้ชิดคุ้นเคยข้อมูลที่เกี่ยวข้องกับความเข้าใจ ตลอดจนข้อมูลที่ไม่เกี่ยวข้องกับผลประโยชน์หรือเรื่องที่ควรปกปิดจะมีความน่าเชื่อถือ หรือน้ำหนักสูงกว่า

ตัวอย่างเช่น ข้อมูลที่ได้รับจากคำถามเรื่องจำนวนที่ดินที่ใช้ในการเพาะปลูกจะมีความถูกต้องหรือน่าเชื่อถือน้อยกว่าข้อมูลจากคำถามเรื่องจำนวนที่ดินถือครอง (กรณีนี้ คือกรณีของความจำหรือการคำนวณ) ข้อมูลจากคำถามเรื่องรายได้ของครอบครัวจะมีความถูกต้องหรือน่าเชื่อถือน้อยกว่าข้อมูลจากคำถามเรื่องเงินเดือนของผู้ตอบเอง (กรณีนี้คือกรณีที่เกี่ยวกับผลประโยชน์)

3. ความน่าเชื่อถือของชุดข้อมูล (Reliability of a Specific Record)

กรณีนี้เป็นกรณีของข้อตกลงของ LSC กล่าวคือข้อมูลรายการใดที่ไม่สามารถตรวจสอบได้ด้วยวิธี Gross Check และวิธีตรวจสอบความแนบนัยได้หรือตรวจด้วยวิธีทั้งสองดังกล่าวแล้วไม่อาจหาข้อยุติได้ กรณีนี้ควรให้น้ำหนักรายการข้อมูลนั้นต่ำกว่ารายการอื่น

4. ชนิดของน้ำหนักใน LSC มี 2 ชนิดคือ w-weight เป็นน้ำหนักที่กำหนดให้แก่รายการข้อมูลและ u-weight เป็นน้ำหนักที่กำหนดให้แก่สมการแนบนัย น้ำหนัก w ควรมากกว่าน้ำหนัก u

ค. การตัดสินใจใช้ข้อมูลทดแทน

เมื่อวิเคราะห์ได้เวคเตอร์ $\tilde{X}$ โดยวิธี LSC แล้ว ให้ใช้ข้อมูลทดแทน $\tilde{X}$ แทนข้อมูล Y เดิมทุกประการ โดยถือว่าข้อมูลทดแทน $\tilde{X}$ เป็นข้อมูลที่ถูกต้อง

13.2 Principal Component Method (PC)

ก. เหตุผลและความจำเป็น

ในการบรรณาธิกรณ์ข้อมูลนั้น โดยปกติเราจำเป็นต้องกระทำกับข้อมูลทุกรายการ หลังจากการตรวจสอบแบบสอบถามครบทั้ง n ชุด เราสามารถสรุปได้ทันทีว่า รายการข้อมูลใดมีความน่าเชื่อถือรายการข้อมูลใดไม่น่าเชื่อถือและต้องหาทางแก้ไขรายการนั้นต่อไป

วิธีการที่จะใช้ในการแก้ไขก็คือ เมื่อตรวจสอบพบว่ารายการข้อมูลใดมีความบกพร่องอยู่เช่น ขาดความแนบนัย ไม่สมเหตุสมผลหรือข้อมูลสูญหาย (โดยปกติจะมีมากกว่า 1 รายการเสมอ) ให้ผู้ตรวจแก้ดำเนินการตรวจสอบดูว่า รายการข้อมูลที่บกพร่องนั้นจะสามารถแก้ไขโดยอาศัยข้อสนเทศจากข้อมูลอื่น ๆ รายการใดบ้าง สมมุติว่าจากการตรวจสอบพบว่ารายการข้อมูลที่พบว่าบกพร่องนั้นสามารถได้โดยอาศัยข้อสนเทศจากข้อมูลรายอื่น ๆ k รายการ ปัญหาที่จะต้องพิจารณากันต่อไปก็คือ จะนำข้อสนเทศเหล่านั้นมาใช้ประโยชน์ได้อย่างไร ทางออกสำหรับปัญหานี้จึงมีอยู่ 2 นัยคือ

(1) อาศัยวิธีการของสมการถดถอย (Regression Analysis) โดยถือว่าข้อมูลที่พบว่าบกพร่องเป็นค่าของตัวแปรตาม ส่วนข้อมูลจากรายการอื่น ๆ ที่เกี่ยวข้องกับรายการที่บกพร่องดังกล่าว เช่นช่วยอธิบายความผันผวนเปลี่ยนแปลงในค่าของรายการนั้นเป็นค่าของตัวแปรอิสระ

ให้ Y = รายการข้อมูลที่บกพร่อง

$X_i's$; i = 1, 2, ..., k คือรายการข้อมูลที่เกี่ยวข้องหรือเป็นรายการที่อธิบายความผันผวนใน Y ได้

ดังนั้นแบบจำลองที่พึงใช้ก็คือ

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + u \quad \text{เมื่อ } u \text{ คือ stochastic error}$$

โดยมีข้อตกลงว่า $E(u) = 0,\ V(u) = \sigma^2,\ E(u_i u_j) = \begin{cases} 0 & ; i \neq j \\ \sigma^2 & ; i = j \end{cases}$

และ $X_i's$ ต้องเป็นอิสระต่อกัน หรือเกี่ยวข้องกันในอัตราค่อนข้างต่ำกล่าวคือ X ต่างก็ทำหน้าที่อธิบายความผันผวนในค่าของ Y โดยที่ X's ต้องไม่สัมพันธ์กันเองจะเกิดปัญหาที่สำคัญคือแยกไม่ออกว่า X ตัวใดแน่ที่มีอิทธิพลต่อ Y และที่สำคัญก็คือค่าประมาณของ β's จะเป็น indeterminate form ($\beta_i = \frac{0}{0}$) และมีความน่าเชื่อถือต่ำ ($V(\beta_i) \to \infty$)

(2) อาศัยวิธีการ PC

วิธี PC คือกรณีเฉพาะของวิธีวิเคราะห์ Factor Analysis กล่าวคือภายหลังจากตรวจสอบพบว่ารายการข้อมูลที่บกพร่อง (Y) สามารถอ้างอิงได้โดยอาศัยข้อมูลสนเทศจากรายการข้อมูลใด (X's) แล้ว ให้ตรวจสอบดูว่า X's มีความเกี่ยวข้องกันอยู่ในอัตราใด โดยอาศัยวิธีวัดค่าสหสัมพันธ์คือ $r_{X_i X_j}$; i, j = 1, 2, ..., k แล้วสร้างตัวแปรอิสระขึ้นมาใหม่โดยอาศัยสหสัมพันธ์เหล่านี้ในลักษณะที่ตัวแปรตัวใหม่เกิดขึ้นจากการอาศัยประโยชน์ของความสัมพันธ์ระหว่าง X's นั่นคือ $Z_i \sum_j^k a_{ij} X_j$: i = 1, 2, ..., k และตัวแปรตัวใหม่เหล่านี้จะถูกสร้างขึ้นมาหลายตัวได้ตามความจำเป็นและความเหมาะสมซึ่ง Z's จะเป็นอิสระต่อกันเสมอ และ Z_1 จะมีความสำคัญสูงกว่า Z_2 Z_2 มีความสำคัญสูงกว่า Z_3 เป็นดังนี้เรื่อย ๆ ไป คำว่ามีความสำคัญสูงกว่า หมายความว่าเป็นตัวแปรที่ดึงดูดเอาอิทธิพลร่วมระหว่าง X's มาไว้ในตัวได้มากกว่า โดยนัยนี้ วิธี PC จึงมีข้อจำกัดน้อยกว่าวิธีวิเคราะห์ความถดถอย เพราะไม่ต้องสนใจว่า X's จะเป็นอิสระต่อกันหรือไม่ และด้วยเหตุที่โดยปกติแล้วตัวแปร X's มักมีส่วนเกี่ยวข้องเชื่อมโยงถึงกันอยู่เสมอ (Intercorrelate) การสร้างข้อมูลทดแทนด้วยวิธี PC จึงต้องตามเหตุผลและสถานะการณ์ได้มากกว่า

ข. ความหมายของ PC

ให้ Y = f(X's) โดยที่ Y คือรายการที่ข้อมูลที่พบว่าบกพร่อง X's คือรายการข้อมูลที่ใช้เป็นแหล่งอ้างอิงของ y

ดังนั้น PC จึงหมายถึงเซทของตัวแปรตัวใหม่คือ Z_i ; i = 1, 2, ..., k ที่เกิดขึ้นจากการประกอบกัน (Linear Combination) ของตัวแปร X's โดยที่

$$Z_i = \Sigma_j^k a_{ij} X_j \quad ; i = 1, 2, \ldots, k$$

เรียก Z_i ว่า PC ที่ i และสัมประสิทธิ์ a_{ij} เรียกว่า loadings โดยที่ a_{ij} เป็นสัมประสิทธิ์ของ X_j ที่ถูกสร้างขึ้นหรือเลือกขึ้นมาในลักษณะที่ทำให้ PC Z_i สอดคล้องกับคุณสมบัติ 2 ประการคือ

(1) PC เป็นอิสระต่อกัน (Uncorrelated หรือ Orthogonal)

(2) PC ตัวแรกจะดึงดูดเอาอิทธิพลร่วมระหว่าง X's ไว้มากที่สุด PC ตัวลำดับถัดไปจะดึงดูดไว้ได้น้อยลงตามลำดับ

ค. วิธีดำเนินการสร้าง PC

การสร้าง PC ให้สอดคล้องกับคุณสมบัติของ loadings 2 ประการข้างต้นมีขั้นตอนดำเนินการกว้าง ๆ ขั้นตอนดังนี้

ขั้นที่ 1 กะประมาณค่าของ loadings $i, j = 1, 2, \ldots, k$ ซึ่งค่า จะช่วยแปลงรูป X's ให้เป็นตัวแปรอิสระ Z's (Orthogonal Artificial Variable) และตรวจสอบนัยสำคัญของ เพื่อดูว่า Z แต่ละตัวควรเกิดขึ้นจากมาประกอบกันของ X's กี่ตัว

ขั้นที่ 2 ตัดสินใจคัดเลือก PC ว่าควรใช้ PC ตัวใด และควรจะคงจำนวน PC ไว้กี่ตัวจึงเพียงพอในการอธิบายความผันผวนของ Y โดยปกติจำนวน PC จะน้อยกว่าจำนวน X's กล่าวคือจำนวนสูงสุดของ PC จะไม่เกินจำนวน X's

ขั้นที่ 3 วิเคราะห์สมการถดถอย

$$Y = \gamma_0 + \sum_{i=1}^{r} r_i Z_i + u \; ; \; r \leq k; \; Z_i = \sum_{j=1}^{k} \hat{a}_{ij} Z_j : i = 1, 2, \ldots, k$$

เพื่อกะประมาณค่าของพารามิเตอร์ γ_i ; $i = 0, 1, 2, \ldots, r$ การกะประมาณให้ใช้วิธี Ordinary Least Square (OLS) ทั้งนี้ถือข้อตกลงเกี่ยวกับ Stochastic error เช่นเดียวกับข้อตกลงของ u ในการวิเคราะห์ความถดถอย

ได้สมการ $\hat{Y} = \hat{\gamma}_0 + \sum_{i}^{r} \hat{\gamma}_i Z_i \quad ; \; r \leq k$

แล้วกะประมาณค่าของ Y โดยอาศัยค่า X's เดิม จากสมการถดถอยข้างต้น

ขั้นที่ 4 กะประมาณค่าของ Y ด้วยช่วงเชื่อมั่น $(1-\alpha)$ 100% ได้ แล้วพิจารณาตัดสินใจเกี่ยวกับความถูกต้องของข้อมูล Y เดิมคือ

ก. ถ้า $\hat{Y}_L \leq Y \leq \hat{Y}_U$ ให้ถือว่าข้อมูล Y เพิ่มถูกต้อง

ข. ถ้า $Y \geq \hat{Y}_U$ ให้ใช้ $\hat{Y}_U$ เป็นข้อมูลทดแทนของ Y

ค. ถ้า $Y < \hat{Y}_L$ ให้ใช้ $\hat{Y}_L$ เป็นข้อมูลทดแทนของ Y

ค.1 การประมาณค่าของ loading a_{ij} ของ Z_i

การประมาณค่าของสัมพันธ์ประสิทธิ์ a_{ij} ให้ดำเนินการดังนี้

(1) หาสหสัมพันธ์ (Simple Correlation) ระหว่าง X's คือ $R_{x_i x_j}$;

$i \neq j = 1, 2, \ldots, k$ ($r_{x_i x_i} = 1$ เมื่อ i=j เมื่อ $r_{x_i x_j} = \frac{\Sigma_s^n x_{is} x_{js}}{\sqrt{\Sigma_s^n x_{is}^2 \Sigma_s^n x_{js}^2}}$ แล้วจัดเรียงไว้ใน

ตารางเรียกว่าตารางสหพันธ์ (Correlation Table) หรือแมตริกซ์สหพันธ์ (Correlation Matrix) แมตริกซ์ดังกล่าวจะเป็น Symmetric Matrix เพราะ $r_{x_i x_j} = r_{x_j x_i}$ สำหรับทุกค่าของ $i \neq j = 1, 2, \ldots, k$ และสมาชิกในแนวทะแยง (Main Diagonol) จะเท่ากับ 1 เสมอเพราะ $r_{x_i x_j} = r_{x_j x_i} = 1$ สำหรับทุกค่าของ i, j ที่ $i = j = 1, 2, \ldots, k$ ดังนี้

	X_1	X_2	X_3	X_4	. . .	X_k
X_1	1	$r_{x_1x_2}$	$r_{x_1x_3}$	$r_{x_1x_4}$	. . .	$r_{x_1x_k}$
X_2	$r_{x_2x_1}$	1	$r_{x_2x_3}$	$r_{x_2x_4}$	. . .	$r_{x_2x_k}$
X_3	$r_{x_3x_1}$	$r_{x_3x_2}$	1	$r_{x_3x_4}$	. . .	$r_{x_3x_k}$
X_4	$r_{x_4x_1}$	$r_{x_4x_1}$	$r_{x_4x_3}$	1	. . .	$r_{x_4x_k}$
.	.	.	.	.	. . .	.
.	.	.	.	.	. . .	.
.	.	.	.	.	. . .	.
X_k	$r_{x_kx_1}$	$r_{x_kx_2}$	$r_{x_kx_3}$	$r_{x_kx_4}$	. . .	1
c_j	$\sum_i^k r_{x_ix_1}$	$\sum_i^k r_{x_ix_2}$	$\sum_i^k r_{x_ix_3}$	$\sum_i^k r_{x_ix_4}$	. . .	$\sum_i^k r_{x_ix_k}$
loadings	l_1	l_2	l_3	l_4	. . .	l_k

1. $x_{is} = X_{is} - \bar{X}_i$, $x_{js} = X_{js} - \bar{X}_j$

(2) รวมค่าของสหสัมพันธ์ในทุก ๆ สดมภ์ (หรือจะใช้แถวแทนสดมภ์ก็ได้) ดังนั้น ผลรวมของสหสัมพันธ์ในสดมภ์ที่ j คือ $c_j = \sum_{i}^{k} r_{X_iX_j}$; $j = 1, 2, \ldots, k$

(3) หายอดรวมทั้งหมดของสหพันธ์ (Grand Total) แล้วถอดรากที่สองนั่นคือ $\sqrt{GT} = (\sum_{j}^{K} \sum_{i}^{R} r_{X_iX_j})^{1/2}$

(4) คำนวณหาค่าประมาณของ loading a_{ij} ของ Z_1 (PC ตัวที่ 1) ด้วยการนำ $\sqrt{GT}$ ไปหารผลรวม c_j

นั่นคือ $\hat{\ell}_i = \hat{a}_{ij} = \dfrac{c_j}{\sqrt{GT}}$; $j = 1, 2, \ldots, k$

หรือ $\hat{\ell}_i = \hat{a}_{ij} = (\sum_{j}^{k} r_{X_iX_j}) / (\sum_{i}^{k} \sum_{j}^{k} r_{X_iX_j})^{1/2}$; $j = 1, 2, \ldots, k$

ดังนั้น $Z_1 = \hat{\ell}_1 X_1 + \hat{\ell}_2 X_2 + \ldots + \hat{\ell}_k X_k$ หรือ

$= \hat{a}_{11} X_1 + \hat{a}_{12} X_2 + \ldots + \hat{a}_{1k} X_k$

ข้อสังเกต 1. Subscript 1 ของ $\hat{a}_{ij}$ หมายถึง PC ตัวที่ 1

2. ขอให้สังเกตว่า $\hat{\ell}_i$ ก็คือสัมประสิทธิ์สหพันธ์รูปหนึ่ง

(5) หา latent root (หรือ eigen value หรือ Charaxcteritic root) ของ PC ตัวที่ 1 โดยที่ latent root λ คือผลรวมกำลังสอง (Sum of Square, SS) ของ loadings กล่าวคือ latent root ของ PC ตัวที่ m คือ

$$\lambda_m = \sum_{j}^{k} \hat{a}_{mj}^2 \ ; \ m \le k$$

ดังนั้น latent root ของ Z_1 คือ

$$\lambda_1 = \sum_{j}^{k} \hat{a}_{1j}^2$$

ด้วยเหตุที่ผลรวมของ latent root ของ PC ทุกตัวมีค่าเท่ากับจำนวนสูงสุดของ FC (หรือจำนวน X's กล่าวคือ $\sum_{m=1}^{k} \lambda_m = k$)

ดังนั้น latent root ของ PC ตัวใดจึงเป็นค่าแสดงความสำคัญของ PC ตัวนั้น เพราะ latent root จะแสดงอัตราความผันแปรรวม (Total Variation) ที่ PC ตัวนั้นดึงดูดออกมาจากกลุ่มของ X's ปริมาณดังกล่าวสามารถวัดเป็นปริมาณของอัตราความผันแปรรวม (Percentage of Total Variation) ได้ดังนี้

$$\text{อัตราของความผันแปรที่ } Z_m \text{ ดึงดูดเอาไว้ได้} = \frac{\lambda_m}{k} \times 100\% \; ; m = 1, 2, \ldots k$$

โดยนัยดังกล่าว Z_1 จึงมี latent root สูงกว่า Z_2 แสดงว่า Z_1 มีความสำคัญหรือดึงดูดเอาความผันผวนหรือมีอิทธิพลร่วมของ X's ไว้ได้มากที่สุด Z_2 มีความสำคัญรองลงไป เป็นเช่นนี้เรื่อย ๆ ไป หรือนัยหนึ่ง $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_k$

(6) จากขั้นดำเนินการทั้ง 5 ขั้น ทำให้ได้ Z_1 พร้อมทั้งปริมาณอิทธิพลร่วม X's ดูดซับเอาไว้ได้คือ λ_1 ดังนี้

$$Z_1 = \hat{a}_{11} X_1 + \hat{a}_{23} X_2 + \ldots + \hat{a}_{1k} X_k \; ; \lambda_1 \; ; \text{ความสำคัญ} = \frac{\lambda_1}{k} \times 100\% \text{ ขั้น}$$

ต่อไปดำเนินการหา Z_2

การหา PC ตัวที่ 2 หรือ Z_2 ให้ดำเนินการเป็นขั้น ๆ ดังนี้

(1) สร้างตาราง Residual Correlation หรือ Residual Correlation Matirx โดยหักลบอิทธิพลของ Z_1 ออกจากตารางสหพันธ์ที่ใช้หา Z_1 ค่าสหพันธ์ผลลัพธ์เรียกว่า Residual Correlation $r^*_{r_i x_j}$ โดยที่ $r^*_{r_i x_j} = r_{x_i x_j} - \hat{a}_{1i} \hat{a}_{1j} \; ; \; i, j = 1, 2, \ldots, k$

ทั้งนี้เพราะ $\hat{a}_{ij}$ คือสัมประสิทธิ์สหสัมพันธ์รูปหนึ่ง

ตัวอย่างเช่น เดิม $r_{x_1 x_2} = r_{x_2 x_1} = .40, \hat{a}_{11} = .30, \hat{a}_{12} = .50$ ดังนั้น Residual Correlation

$$r^*_{x_1 x_2} - r_{x_1 x_2} = \hat{a}_{11} \hat{a}_{12} = 40. - (30 - 50) = 25$$

ดังนั้น Residual Correlation Table สำหรับ Z_2 จึงปรากฏดังนี้

	X_1	X_2	....	X_k
X_1	$r^*_{X_1X_1} = 1-\hat{a}^2_{11}$	$r^*_{X_1X_2} = r_{X_1X_2} = r_{X_1X_2} - \hat{a}_{11}\hat{a}_{12}$	...	$r^*_{X_1X_k} = r_{X_1X_k} - \hat{a}_{11}\hat{a}_{1k}$
X_2	$r^*_{X_2X_1} = r_{X_2X_1} - \hat{a}_{12}a_{11}$	$r^*_{X_2X_2} = 1 - \hat{a}^2_{12}$	...	$r^*_{X_2X_k} = r_{X_2X_k} - \hat{a}_{12}\hat{a}_{1k}$
X_3	$r^*_{X_3X_1} = r_{X_3X_1} - \hat{a}_{13}\hat{a}_{11}$	$r^*_{X_3X_2} = r_{X_3X_2} - \hat{a}_{13}\hat{a}_{12}$	...	$r^*_{X_3X_k} = r_{X_3X_k} - \hat{a}_{13}\hat{a}_{1k}$
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
Xk	$r^*_{X_kX_1} = r_{X_kX_1} - \hat{a}_{1k}\hat{a}_{11}$	$r^*_{X_kX_2} = r^*_{X_kX_2} - \hat{a}_{1k}\hat{a}_{2k}$	...	$r^*_{X_kX_k} = 1 - \hat{a}^2k$
รวม	$\sum_{i}^{k} r^*_{X_iX_1}$	$\sum^{k} r^*_{X_iX_2}$	...	$\sum r_{X_iX_k}$
loadings	$\hat{a}_{21}$	$\hat{a}_{22}$	...	$\hat{a}_{2k}$

ขอให้สังเกตว่า $r^*_{X_iX_j} \neq 1$ เมื่อ $i=j$

(2) จาก Residual Correlation Matrix Loading ได้

$$\hat{a}_{2j} = \left(\sum_{i}^{k} r^*_{X_iX_j}\right) / \left(\sum_{i}^{k}\sum_{i}^{k} r^*_{X_iX_j}\right)^{1/2} \quad ; \quad j = 1, 2, \ldots, k$$

(3) คำนวณหา latent root ของ z_2 ได้

$$\lambda_2 = \sum_{j}^{k} \hat{a}^2_{2j}$$

พร้อมทั้งปริมาณความผันแปรในอิทธิพลของ X's ที่ Z_2 ดึงดูดไว้ภายหลังจากที่ Z_i ได้ดึงดูดไปแล้วนั่นคือ

$$\text{อัตราความผันแปรที่ } Z_2 \text{ ดึงดูดไว้ได้} = \frac{\lambda_2}{k} \times 100\%$$

(4) คำนวณหา Z_2 จะพบว่า

$$Z_2 = \hat{a}_{21} X_1 + \hat{a}_{22} X_2 + \ldots + \hat{a}_{2k} X_k \; ; \text{ ความสำคัญ } = \frac{\lambda_2}{k} \times 100\%$$

การหา Z_3 ให้หาโดยการสร้าง Residual Correlation Matrix ขึ้นใหม่จากตาราง Residual Correlation ของ Z_2 แล้วดำเนินการแบบเดิมจะได้

$$Z_3 = \hat{a}_{31} X_1 + \hat{a}_{32} X_2 + \ldots + \hat{a}_{3k} X_k \; ; \text{ ความสำคัญ } = \frac{\lambda_2}{k} \times 100\%$$

โดยที่

$$\hat{a}_{3j} = \left(\sum_i^k r^{**}_{X_iX_j}\right) / \left(\sum_i^k \sum_j^k r^{**}_{Z_iZ_j}\right)^{1/2} \quad ; \quad j = 1, 2, \ldots, k$$

$$r^{**}_{X_iX_j} = r^{*}_{X_iX_j} - \hat{a}_{2i}\,\hat{a}_{2j} \quad ; \quad i, j = 1, 2, \ldots, k$$

$$\lambda_3 = \sum_j^k \hat{a}^2_{3j}$$

สำหรับ Z_4, Z_5, ..., Z_k ก็ดำเนินการแบบเดียวกันนี้ กล่าวคือสร้าง Residual Correlation Matrix ขึ้นมาโดยอาศัย Residual Correlation Matrix ของ PC ตัวก่อนหรือลำดับก่อน

ค.2 การตัดสินใจคัดเลือกองค์ประกอบของ Z_i

การพิจารณาคัดเลือกองค์ประกอบของ Z_i หมายถึงการตัดสินใจว่า Z_i ควรประกอบไปด้วยตัวแปร X's กี่ตัวจึงจะถือว่าเหมาะสมหรือเพียงพอ เกณฑ์การตัดสินใจก็คือนัยสำคัญของ a_{ij} ถ้า a_{ij} มีนัยสำคัญ แสดงว่า X_j ควรคงอยู่หรือควรถือว่า X_j เป็นองค์ประกอบของ Z_i หรือ X_i ควรเป็นปัจจัยที่มีอิทธิพลร่วมให้เกิดตัวแปร Z_i

ดังนั้น จาก PC ใด ๆ คือ

$$Z_i = \sum_j^k \hat{a}_{ij} X_j \quad ; \quad i = 1, 2, \ldots, k$$

หรือ $Z_i = \hat{a}_{i1} X_1 + \hat{a}_{i2} X_2 + \ldots + \hat{a}_{ik} X_k$

สิ่งที่ต้องการคือ a_{ij} ; $j = 1, 2, \ldots, k$ มีนัยสำคัญ ระดับนัยสำคัญ α หรือไม่ นั่นคือเราต้องทำการทดสอบสมมุติฐานต่อไปนี้

$H_0 : a_{ij} = 0 =$ vs $H_1 : a_{ij} \neq 0 ; j = 1, 2, \ldots, k$ สำหรับ i ใด ๆ โดยที่ $i = 1, 2, \ldots, k$

เช่น ต้องการทดสอบว่า Z_1 ควรอาศัยอิทธิพลร่วมของ X ตัวใดบ้าง ดังนั้นจากโครงสร้าง

$$Z_1 = \sum_{j}^{k} a_{ij} X_j \qquad \text{หรือ}$$

$$Z_1 = a_{11} X_1 + a_{12} X_2 + \ldots + a_{1k} X_k$$

เราต้องทำการทดสอบสมมุติฐานต่อไปนี้

$$H_0 : a_{1j} = 0 \text{ vs } H_1 : a_{1j} \neq 0 : j = 1, 2, \ldots, k$$

การตรวจสอบนัยสำคัญของ a_{ij} ; $j = 1, 2, \ldots, k$ สำหรับ fixed i $i = 1, 2, \ldots, k$ นั้นมีวิธีการต่าง ๆ หลายวิธีดังนี้

(1) Empirical Test

การทดสอบวิธีนี้เป็นวิธีที่ค่อนข้างหยาบแต่ใช้ได้ผลดีสำหรับกรณีขนาดตัวอย่าง n มีค่าน้อยเท่ากับ 50 หน่วย ($n \leq 50$) กล่าวคือ loadings a_{ij} ใดจะมีนัยสำคัญก็ต่อเมื่อ $|a_{ij}| > .30$ หรือ $-.30 > a_{ij} > .30$ หรือนัยหนึ่งปฏิเสธสมมุติฐานหนัก $H_0 : a_{ij} = 0$ และยอมรับสมมุติฐานรอง $H_1 : a_{ij} \neq 0$ เมื่อ $|a_{ij}| > .30$ หรือ a_{ij} มากกว่า $\pm .30$

(2) Standard error Test

เนื่องจาก $\hat{a}_{ij} = (\sum_{j}^{k} r_{X_iX_j}) / (\sum_{i}^{k}\sum_{j}^{k} r_{X_iX_j})^{1/2}$; $:i = 1, 2, \ldots, k$

คือรูปหนึ่งของสัมประสิทธิ์สหสัมพันธ์ การทดสอบ $H_0 : a_{ij} = 0$ vs $H_1 : a_{ij} \neq 0 : j = 1, 2, 3, \ldots, k$ สำหรับ fixed i $i = 1, 2, 3, \ldots, k$ จึงทดสอบเช่นเดียวกับการทดสอบสหสัมพันธ์ r (Pearson's Product Monent) กล่าวคือใช้ตารางค่าวิกฤติของ r ตรวจสอบนัยสำคัญของ a_{ij} ค่าวิกฤติของ r ก็คือส่วนเบี่ยงเบนมาตรฐานของ r ซึ่งจะใช้เป็นส่วนเบี่ยงเบน

มาตรฐานของ a_{ij} ต่อไปนั่นเอง

การตรวจสอบนัยสำคัญของ a_{ij} ให้กระทำดังนี้

จาก $H_0 : a_{ij} = 0$ vs $H_1 ; a_{ij} \neq 0 ; j = 1, 2, \ldots, k$ สำหรับ fixed i $i = 1, 2, \ldots, k$

เราจะปฏิเสธสมมุติฐานหลักและยอมรับสมมุติฐานรองเมื่อ $|a_{ij}| > r_{n,\alpha}$ เมื่อ $r_{n,\alpha}$ คือค่าวิกฤติ ณ. ระดับนัยสำคัญ α องศาความเป็นอิสระ n ซึ่งเป็นค่าในตารางวิกฤติของ

ขนาดตัวอย่าง	ค่าวิกฤติสหสัมพันธ์ ณ. ระดับนัยสำคัญ	
	5 %	1 %
5	.755	.875
10	.076	.714
15	.483	.605
20	.425	.538
25	.380	.488
30	.338	.440
35	.320	.417
40	.300	.394
45	.280	.370
50	.262	.346
60	.248	.328
70	.233	.308
80	.220	.290
90	.206	.272
100	.190	.255
150	.158	.209
200	.138	.182
250	.125	.163
500	.086	.115

(2) Burt-Bank Test (B-B)

วิธีทดสอบที่ (1) และ (2) เป็นวิธีที่ค่อนข้างหยาบทั้งนี้เพราะมิได้คำนึงถึงจำนวนตัวแปร X's ที่ร่วมเป็นองค์ประกอบร่วมของ Z และลำดับความสำคัญของ Z เอง ทำให้การตรวจจับนัยสำคัญของวิธีทดสอบดังกล่าวมีช่องทางให้คัดค้านได้ วิธีทดสอบของ B-B จะแก้จุดอ่อนข้อนี้ได้โดยนำทั้งจำนวนตัวแปร X และลำดับความสำคัญของ Z มาร่วมพัฒนาตัวทดสอบเพื่อตรวจสอบนัยสำคัญของ loadings โดยนำค่าทั้งสองดังกล่าวมาปรับปรุงค่าส่วนเบี่ยงเบนมาตรฐานของ r ให้รัดกุมเสียก่อน ก่อนที่จะนำค่าดังกล่าวไปใช้เป็นค่าวิกฤติ

ให้ k = จำนวนตัวแปร X

m = Subscript ของ Z ซึ่งแสดงลำดับความสำคัญของ Z (อย่าลืมว่า Z_1 สำคัญกว่า Z_2 Z_2 สำคัญกว่า $Z_3, \ldots, Z_{k-1}$ สำคัญว่า Z_k)

$s(r_{x_i x_j})$ = ค่าวิกฤต (ส่วนเบี่ยงเบนมาตรฐาน) ของ r ณ. ระดับสำคัญ α อิสระ n (ขนาดตัวอย่าง) ซึ่งในที่นี้ $s(r_{x_i x_j})$ ก็คือค่าวิกฤติของ a_{ij} หรือ $S(r_{x_i x_j}) = r_{n,\alpha}$ ดังนั้นค่าวิกฤตของ a_{ij} ที่ปรับปรุงแล้วคือ $s(a_{mj})$ โดยที่

$$S(a_{mj}) = s(r_{x_i x_j})\sqrt{\frac{k}{k+1-m}}$$

$$= r_{n,\alpha}\sqrt{\frac{k}{k+1-m}}$$

นั่นคือวิกฤติของ loading ตัวที่ $j : j = 1, 2, \ldots, k$ ของ PC ตัวที่ m คือ

$$s(a_{mj}) = r_{n,\alpha}\sqrt{\frac{k}{k+1-m}} \; ; \; j = 1, 2, \ldots, k$$

ซึ่งเราจะปฏิเสธสมมุติฐานหลัก $H_0 : a_{ij} = 0 ; j = 1, 2, \ldots, k ; i = m$ และยอมรับสมมุติฐานรอง $H_1 ; a_{ij} \neq 0 ; j = 1, 2, \ldots, k ; i = m$ เมื่อ

$|a_{mj}| > s(a_{mj}) : i = 1, 2, \ldots, k$ สำหรับ fixed i ; i = m

ค.3. การพิจารณาคัดเลือกจำนวน PC ที่จะเหลือไว้ในการวิเคราะห์

โดยปกติจำนวน PC จะอยู่ไม่เกินจำนวนตัวแปร X และเหตุที่สมการ $Y = f(X's)$ นั้นควรจะคัดเลือกตัวแปร Z ไว้เฉพาะตัวที่มีความสำคัญจริง ๆ หรือตัวที่ดึงดูดเอาอิทธิพลร่วมระหว่าง X ไว้ได้มากจริง ๆ เท่านั้น วิธีการตัดสินใจคัดเลือกมีอยู่หลายวิธี จะกล่าวถึงเพียง 2 วิธี ดังนี้

(1) Kaiser's Criterion

วิธีคัดเลือกของไกเซอร์พิจารณาเลือก PC เฉพาะตัวที่ให้ latent root มากกว่า 1 เท่านั้นตัวที่ให้ latent root น้อยกว่า 1 จะถูกตัดทิ้ง

นั่นคือคัดเฉพาะ Z_m ที่ $\lambda_m > 1$ ไว้ในการวิเคราะห์ แต่โดยเหตุที่ $\lambda_1 > \lambda_2 > \ldots \lambda_k$ ดังนั้นเราจะหยุดวิเคราะห์ PC ตัว $(r+1)$ ถ้าพบว่า $\lambda_r < 1$ ซึ่งจะช่วยประหยัดเวลาและแรงงานลงไปได้มากทั้งนี้เพราะ $\lambda_{T+1}, \lambda_{T+2}, \ldots, \lambda_k$ จะมีค่าน้อยกว่า 1 อย่างแน่นอน

กฎเกณฑ์ของไกเซอร์เหมาะที่จะใช้เป็นหลักในการคัดเลือก PC หรือมีความน่าเชื่อถือได้เฉพาะเมื่อ $20 < k < 50$ หรือมีจำนวนตัวแปร X อยู่ระหว่าง 20-50 ตัว ถ้า $k > 50$ จะมีแนวโน้มให้เหลือ PC ไว้มากเกินไป ขณะที่ $k < 20$ จะมีแนวโน้มให้เหลือ PC ไว้น้อยเกินไป

(2) Bartlett's Criterion

กฎเกณฑ์ของบาร์ตเลต์มีลักษณะการคัดเลือกคล้ายวิธีของไกเซอร์ในแง่ที่พิจารณาความสำคัญของ PC โดยพิจารณาจาก latent root โดยมีหลักเกณฑ์การคงจำนวน PC ไว้เช่นเดียวกัน ต่างกับวิธีของไกเซอร์ที่บาร์ตเลต์มีวิธีคัดเลือก PC ตัวต่อไปได้รับกุมกว่า ดังนี้

สมมุติว่า $\lambda_1, \lambda_2, \ldots, \lambda_r$; $r < k$ มีค่าสูงและแตกต่างกัน ซึ่งทำให้เชื่อได้ว่า $Z_1, Z_2, \ldots, Z_k$ ควรจะคงไว้ในสมการ $Y = f(Z's)$ ปัญหาก็คือ PC ที่เหลืออยู่อีก k-r ดังนั้นควรจะคงไว้ด้วยหรือไม่ ควรจะนำไปเพิ่มในสมการ $Y = f(X's)$ ได้อีกหรือไม่ ทั้งนี้ให้ถือว่า Z_k มีความสำคัญน้อยที่สุดและไม่ควรนำเข้าไปในสมการโครงสร้าง

ด้วยปัญหาดังกล่าว บาร์ตเลต์จึงเสนอให้ทำการทดสอบสมมุติฐานดังนี้

H_0 : PC ที่เหลืออยู่ (k-r) ตัวไม่มีความสำคัญ หรือเสนอเป็นสัญญลักษณ์ได้ดังนี้

$H_0 : \lambda_{T+1} = \lambda_{T+2} = \ldots = \lambda_k$

ขณะที่สมมุติฐานรองคือ

H_1 : ควรเพิ่มจำนวน PC เข้าไปในสมการโครงสร้างได้อีก หรือเสนอเป็นสัญญ-ลักษณ์ได้ ดังนี้

H_1 : λ_i not all equal ; $i = r+1, r+2, \ldots, k$

ตัวสถิติที่ใช้ทดสอบคือ

$$X^{*^2} = n\ln \left\{ (\lambda_{r+1} . \lambda_{r+2}, \ldots \lambda_k)^{-1} \left(\frac{\lambda_{r+1} + \lambda_{r+2} + \ldots + \lambda_k}{k-2} \right)^{k+1} \right\}$$

โดยที่ $X^{*^2} \sim X^2_{\gamma,\alpha}$ เมื่อ $\gamma = \frac{1}{2}(k-r-1)(k-r+2)$

และปฏิเสธสมมุติฐานหลักเมื่อ $X^{*^2} > X^2_{\gamma,\alpha}$ ซึ่งหมายความว่าควรจะเพิ่ม PC เข้าไปในสมการโครงสร้างได้ การเพิ่มจะต้องเพิ่มคราวละตัวแล้วทดสอบสมมุติฐานต่อไปจนกว่าจะยอมรับสมมุติฐานหลักซึ่งหมายความว่า PC ที่เหลืออยู่มีความสำคัญมากไม่ควรนำเพิ่มเข้าไปในสมการ

ค. วิเคราะห์สมการถดถอย $Y = \gamma_0 + \sum_i^r \gamma_i Z_i + u$

การวิเคราะห์สมการถดถอย $Y = \gamma_0 + \sum_i^r \gamma_i Z_i + u$ ยึดถือวิธี OLS ในการประมาณค่าพารามิเตอร์ γ_i ; $i = 0, 1, 2, \ldots, r$

โดยที่ $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_r)^T = (Z^T Z)^{-1} Z^T Y$

$$\hat{\sigma}^2 = \frac{1}{n-r} (Y^T Y - \hat{\gamma}^T Z^T Y)$$

และ $V(\hat{\gamma}) = \hat{\sigma}^2 ((Z^T Z)^{-1}$

ทั้งนี้อาศัยการวิเคราะห์แบบ Linear Hypothesis Model of Full Rank เนื่องจากตัวแปร Z's เป็นอิสระต่อกัน (Uncorrelate หรือ Orthogonal) โดยวิเคาะห์จากแบบจำลอง

$$Y_i = \gamma_0 + \sum_i^r l_i Z_{ij} + u_i \quad ; j = 1, 2, 3, \ldots, n$$

เมื่อ $Z_i = \sum_{j}^{k} \hat{a}_{ij} X_j$; $i = 1, 2, 3, \ldots . r$

หรือจัดในรูปสมการแมตริกซ์

$$Y_{n\times 1} = Z_{n\times(r=1)} \gamma_{(r+1)\times 1} + u_{n\times 1}$$

และด้วยเหตุที่งานวิเคราะห์นี้มีจุดมุ่งหมายเพื่อการพยากรณ์ คือ คาดหมายข้อมูลทดแทน เครื่องมือที่ใช้ในการตัดสินใจเลือก Best fitted model จึงใช้สัมประสิทธิ์แห่งการตัดสินใจ $\bar{R}^2$ (coeffieient of Determination) โดยที่

$$\bar{R}^2 = 1 - \{ (Y^TY - \hat{\gamma}^T Z^T Y)/(n-r)\} / \{(Y^T Y - n\bar{Y}^2)/(n-1)\}$$

นั่นคือ Estimated model ใดให้ $\bar{R}^2$ สูงสุดก็พึงเลือกใช้ model นั้น ทั้งนี้โดยมิต้องคำนึงว่า γ_i ; $i = 1, 2, \ldots, r$ จะมีนัยสำคัญ หรือไม่มากนัก

สมการพยากรณ์ที่ต้องการคือ

$$\begin{array}{lll} \hat{Y} & = & \hat{\gamma}_0 + \hat{\gamma}_1 Z_1 + \hat{\gamma} Z_1 + \hat{\gamma}_3 Z_3 + \cdots . + \hat{\gamma}_r Z_r \\ & & (s_{\hat{\gamma}_0})\ (s_{\hat{\gamma}_1})\ (s_{\hat{\gamma}_2})\ (s_{\hat{\gamma}_3}) \quad \ldots \qquad (s_{\hat{\gamma}_r}) \\ & & \qquad (t_{\hat{\gamma}_1})\ (t_{\hat{\gamma}_2})\ (t_{\hat{\gamma}_3}) \quad \ldots \qquad (t_{\hat{\gamma}_r}) \end{array}$$

โดยที่ ; $i = 1, 2, \ldots, r-1$, $s_{\hat{\gamma}_i} = \sqrt{c_{ii}}$

เมื่อ c_{ii} คือสมาชิกในตำแหน่งที่ $(i+1, i+1)$ ของแมตริกซ์

$\hat{\sigma}^2(Z^TZ)^{-1}$ และ $t_{\hat{\gamma}_i} = \dfrac{\hat{\gamma}_i - \gamma_i}{\sqrt{c_{ii}}} \sim t_{n-1,\, 1-\alpha/2}$ ซึ่งใช้ตรวจสอบนัยสำคัญของ γ_i ; i 1, 2, ..., r หรือสมมุติฐาน $H_0 : \gamma_i = 0$; vs $H_1 : \gamma_i \neq 0$; $i = 1, 2, \ldots, r$ และปฏิเสธสมมุติฐานหลัก $H_0 : \gamma_i = 0$ $i = 1, 2, \ldots . r$ เมื่อ $|t_{\hat{\gamma}_i}| > t_{n-1\ 1-\alpha/2}$

สำหรับข้อตกลงเบื้องต้นเกี่ยวกับ Stochastic error ให้ยึดถือข้อตกลงของ u ใน OLS ทุกประการคือ

$$E(u) = 0,\ V(u) = \sigma^2,\ E(u_i u_j) = \begin{cases} 0 \ ; \ i \neq j \\ \sigma^2 \ ; \ i \neq j \end{cases} \quad , u \sim N(O, \sigma^2)$$

ค.4 การพยากรณ์ค่าของ Y ด้วยช่วงเชื่อมั่น

ให้ Z_{iF} = ค่าพยากรณ์ของ Z_i ใน forecasting period ในที่นี้จะถือว่า $Z_{iF} = Z_i$

n = ขนาดตัวอย่าง ในที่นี้ n คือจำนวนชุดของแบบสอบถามหรือจำนวน record

$\hat{\sigma}^2$ = ค่าประมาณของ σ^2

ดังนั้นค่า ค่าพยากรณ์ของค่าเฉลี่ยของ Y_{F_i} คือ

$$\hat{Y}_{F_i} = \hat{\gamma}_0 + \hat{\gamma}_1 Z_{1F_i} + \hat{\gamma} Z_{2F_i} + \ldots + \hat{\gamma} Z_{rF_i} \ ; \ i = 1, 2, \ldots, n$$

ให้ $Z_F = (1, Z_{1F}, Z_{2F}, \ldots, Z_{rF})$ เมื่อ Z_F เป็นแมตริกซ์ขนาด $n \times (r+1)$

ซึ่งสมาชิกของ Z_F คือ Column Vector ขนาด $n \times 1$ กล่าวคือ

$$1 = (1, 1, 1, \ldots, 1)^T,\ Z_{1F} = (Z_{21}, Z_{22}, Z_{27}, \ldots, Z_{2n})^T$$

$$Z_{2F} = (Z_{31}, Z_{32}, Z_{33}, \ldots, Z_{3n})^T, \ldots, Z_{rF} = (Z_{(r+1)\times 1}, Z_{(r+1)\times 2}, \ldots, Z_{(r+1)\times n})^T$$

ให้ $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_r)^T \qquad Y = (Y_1, Y_2, Y_3, \ldots, Y_n)^T$

ดังนั้น

$$Y = Z\hat{Y} = (1, Z_{1F}, Z_{2F}, \ldots, Z_{rF}) \begin{pmatrix} \hat{\gamma}_0 \\ \gamma_1 \\ \hat{\gamma}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\gamma}_r \end{pmatrix}$$

หรือ

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \hat{Y} \\ . \\ . \\ . \\ \hat{Y}_n \end{pmatrix} \cdot = \begin{pmatrix} 1 & Z_{21} & Z_{31} & \dots & Z_{r1} \\ 1 & Z_{22} & Z_{33} & \dots & Z_{r2} \\ 1 & Z_{23} & Z_{33} & \dots & Z_{r3} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ 1 & Z_{2n} & Z_{3n} & \dots & Z_{rn} \end{pmatrix} \begin{pmatrix} \gamma'_0 \\ \hat{\gamma}_1 \\ \hat{\gamma}_3 \\ . \\ . \\ . \\ \gamma^r \end{pmatrix}$$

และเนื่องจาก $\hat{Y} = (1, Z_{1F}, Z_{2F}, \dots, Z_{rF}) (\hat{\gamma}_0, \hat{\gamma}_2, \hat{\gamma}_2, \dots, \hat{\gamma})^T$ โดยกำหนดให้ Z_{iF} เป็นแวคเตอร์ของค่าคงที่ของ X ซึ่งในที่นี้ถือว่า $Z_{iF} = Z_i$ หมายความว่าถือว่าค่าเดิมของ Z_i เป็นค่าเดียวกับค่าพยากรณ์ของ Z_i ดังนั้น $\hat{Y}$ จึงเป็น linear Combination ของตัวแปรสุ่ม $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_3, \dots, \hat{\gamma}_r$

ดังนั้น $V(\hat{Y}_{Fi}) \quad = V(\hat{\gamma}_0 + \hat{\gamma}_1 Z_{1i} + \hat{\gamma}_2 Z_{2i} + \dots + \hat{\gamma}_r Z_{ri}) \; ; \; i = 1, 2, \dots, n$

หรือ $\quad = \sum_{i=0}^{r+1} \sum_{j=0}^{r+1} V_{ij} Z_i Z_j \; ; \; V_{ij} = \begin{cases} V(\gamma_i) \; ; \; i=j \\ Cov(\gamma_i, g_j) \; ; \; i \neq j \end{cases}$

จะเห็นได้ว่า $V(Y_{Fi})$ เป็น Quadratic form

ดังนั้น

$$V(Y_{Fi}) = (1, Z_{2i}, Z_{3i}, \dots, Z_{ri}) \begin{pmatrix} V(\lambda_0) & Cov(\gamma_0, \gamma_1) & Cov(\gamma_0, \gamma_2) & \dots Cov(\gamma_0, \gamma_r) \\ Cov(g_1, \gamma_0) & V(\gamma_1) & Cov(\gamma_1, \gamma_2) & \dots Cov(\gamma_1, \gamma_r) \\ . & & & . \\ . & & & . \\ . & & & . \\ Cov(\gamma_r, \gamma_0) & Cov(g_r, \gamma_1) & Cov(\gamma_r, \gamma_2) & \dots V(\gamma_r) \end{pmatrix} \begin{pmatrix} 1 \\ Z_{2i} \\ . \\ . \\ . \\ Z_{ri} \end{pmatrix}$$

$i = 1, 2, \dots, n$

$$= (1, Z_{2i}, Z_{3i}, \dots, Z_{ri}) \cdot (Z^T Z)^{-1} \hat{\sigma}^2 \cdot (1, Z_{2i}, Z_{3i}, \dots, Z_{ri})^T \; ; \; i = 1, 2, \dots, n$$

$$= \hat{\sigma}^2 (1, Z_{2i}, Z_{3i}, \dots, Z_{ri}) (Z^T Z)^{-1} (1, Z_{2i}, Z_{3i}, \dots, Z_{ri})^T \; ; \; i = 1, 2, \dots, n$$

โดยที่ $\hat{\sigma}^2 = \frac{1}{n-r} (\gamma^T Y - \hat{\gamma}^T Z^T Y)$

1 Wilk, SS., "Methematical Statistics" (John Wiley & Sons Inc, Tokyo 1962) p. 82

ดังนั้นช่วงพยากรณ์ (1 - α) 100% ของ Y_{F_i} ; i = 1, 2, ..., n คือ

$$\hat{Y}_{F_i} + t_{n-r,\alpha/2} \sqrt{V(\hat{Y}_{F_i})} < Y_{F_i} < \hat{Y}_{F_i} + t_{n-r,1-\alpha/2} \sqrt{V(\hat{Y}_{F_i})}$$

หรือ $$Y_L \leq Y_i \leq Y_U$$

โดยที่ $$\hat{Y}_{F_i} = \hat{\gamma}_0 + \hat{\gamma}_1 Z_{1i} + \hat{\gamma}_2 Z_{2i} + \ldots + \hat{\gamma}_r Z_{ri}$$

$$V(\hat{Y}_{F_i}) = \hat{\sigma}^2 (1, Z_{2i}, Z_{3i}, \ldots, Z_{ri}) (Z^T Z)^{-1} (1, Z_{2i}, Z_{3i}, \ldots, Z_{ri})$$

$$\hat{\sigma}^2 = \frac{1}{n-r} (Y^T Y - \hat{\gamma}^T Z^T Y)$$

ค่าช่วงพยากรณ์ (1-α) 100% นี้เรียกว่าช่วงของข้อมูลทดแทน (Y_L, Y_U)

ค. 5 การตัดสินใจเกี่ยวกับความถูกต้องน่าเชื่อถือของข้อมูลเดิม

ให้ Y_i คือข้อมูลเดิม และ Y_{F_i} คือข้อมูลทดแทนของ Y_i ที่ได้จากวิธีการ PC ดังนั้น ถ้า

(1) $Y_L < Y_i < Y_U$ 0 0 ให้ถือว่าข้อมูลเดิมถูกต้อง

(2) ถ้า $Y_i > Y_U$ ให้ใช้ $Y_U = \hat{Y}_{F_i} + t_{n-r,1-\alpha/2} \sqrt{V(\hat{Y}_{F_i})}$ เป็นข้อมูลทดแทน

(3) ถ้า $Y_i < Y_L$ ให้ใช้ $Y_L = \hat{Y}_{F_i} + t_{n-r,\alpha/2} \sqrt{V(\hat{Y}_{F_i})}$ เป็นข้อมูลทดแทน

1.3.3 วิธีบรรณาธิกรณ์เชิงตรรกวิทยา (Logical Editing) [1]

ก. เหตุผลและความจำเป็น

ในกรณีที่ข้อมูลต้องการบรรณาธิกรณ์และแก้ไขเป็นข้อมูลเชิงคุณภาพ (Qualitive Data) เช่นเพศ สีผิว เชื้อชาติ ศาสนา สถานภาพสมรส ความสัมพันธ์กับหัวหน้าครอบครัว ภูมิลำเนาเดิม ฯลฯ การบรรณาธิกรณ์และการสร้างข้อมูลทดแทนย่อมมีวิธีการแตกต่างไปจากวิธีทั้งสองข้างต้น เพราะวิธี LSC และ PC เหมาะสำหรับใช้กับกรณีของข้อมูลเชิงปริมาณ (Quantitative Data or Metric) ทั้งนี้ด้วยเหตุผลประการสำคัญคือข้อมูลทดแทนที่ได้รับจาก

---

1. P. Fellegi and D. Holt, A Systematic Approach to Automatic Edit and Ilpntation (Jour. Am. Stat. Asso.) P.17

วิธีดังกล่าวไม่สามารถแปลความหมายกลับเป็นเชิงคุณภาพได้ชัดเจนหรือรัดกุมเพียงพอ ด้วยเหตุดังกล่าวจึงได้มีความพยายามนำเอาหลักเกณฑ์ทางตรรกวิทยามาใช้เป็นแนวทางในการบรรณาธิกรณ์และสร้างข้อมูลทดแทนสำหรับกรณีของข้อมูลเชิงคุณภาพดังกล่าว วิธีการนี้เรียกว่า การบรรณาธิกรณ์เชิงตรรกวิทยา (Logical Editing)

หลักการโดยสรุปของการบรรณาธิกรณ์เชิงตรรกวิทยาประกอบด้วยขั้นดำเนินการกว้าง ๆ 4 ขั้นดังนี้คือ

1. สร้างเงื่อนไขการตรวจสอบเรียกว่าเงื่อนไขขัดแย้ง (fail edit, e) ขึ้นโดยพิจารณาจากระหัสคำตอบ (Code) ที่เป็นไปได้ทั้งหมด ของแบบสำรวจ (Record) แล้วประมวล Combination ของระหัสที่ขัดแย้งกับความเป็นจริงขึ้นมาเป็นเซท
2. นำ Record ที่ต้องการมาตรวจสอบตามเงื่อนไข
3. สร้างข้อมูลทดแทนให้แก่รายการข้อมูล (field) ที่ขัดแย้งกับความเป็นจริง หรือสอดคล้องกับเงื่อนไขขัดแย้ง (fail edit)

ข. ตัวอย่าง

เพื่อความสะดวกต่อการศึกษาวิธีการตลอดจนทฤษฎีของวิธิกรณ์เชิงตรรกวิทยา (logical editing) ซึ่งจะกล่าวถึงต่อไป ควรจะได้ทำความเข้าใจในตัวอย่างต่อไปนี้ซึ่งจักเป็นแนวทางให้เข้าใจสัญญลักษณ์และนิยามบางประการที่เกี่ยวข้อง

สมมุติแบบสอบถามชุดหนึ่ง Record ประกอบด้วยรายการข้อมูล (field) 3 รายการ คือเพศ อายุและความสัมพันธ์กับหัวหน้าครอบครัว

จากการพิจารณาตามหลักแห่งเหตุผลและความเป็นไปได้ เราสามารถจำแนกระหัสคำตอบของ

แต่ละรายการข้อมูลได้ดังนี้

เพศ	อายุ	ความสัมพันธ์กับหัวหน้าครอบครัว
ชาย	0-14	ภรรยา
หญิง	15 +	สามี
		ธิดา
		ญาติ
		อื่น ๆ

ให้ A_i = เซทของระหัสคำตอบของรายการข้อมูลที่ i ; $i = 1, 2, 3$

n_i = จำนวนระหัสหรือจำนวนสมาชิกของเซท A_i

ดังนั้น

A_1 = {ชาย, หญิง} ; $n_1 = 2$

A_2 = {0-14 ปี, 15+ ปี} ; $n_2 = 2$

A_3 = {ภรรยา, สามี, บุตร, ธิดา, ญาติ, อื่น ๆ} ; $n_3 = 6$

ดังนั้น คำตอบทั้งหมด (all possible code) ของแบบสอบถาม (Record) ชุดนี้จึงประกอบด้วยคำตอบทั้งนั้น $2 \times 2 \times 6 = 24$ คำตอบคือ

A = {(ชาย, อายุไม่เกิน 14 ปี, เป็นภรรยา), (ชาย, อายุตั้งแต่ 15 ปี ขึ้นไป, เป็นภรรยา), (ชาย, อายุไม่เกิน 14 ปี, เป็นสามี), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นสามี), (ชาย, อายุไม่เกิน 14 ปี, เป็นบุตร), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นธิดา), (ชาย, อายุไม่เกิน 14 ปี, เป็นญาติ), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นญาติ), (ชาย, อายุไม่เกิน 14 ปี, เกี่ยวข้องในฐานะอื่น), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เกี่ยวข้องในฐานะอื่น), (หญิง, อายุไม่เกิน 14 ปี, เป็นภรรยา), (หญิง, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นภรรยา), (หญิง, อายุไม่เกิน 14 ปี, เป็นสามี), (หญิง, อายุตั้งแต่ 15 ปี ขึ้นไป, เป็นสามี), (หญิง, อายุไม่เกิน 14 ปี, เป็นบุตร), (หญิง, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นธิดา), (หญิง, อายุไม่เกิน 14 ปี, เป็นญาติ), (หญิง, อายุไม่เกิน 14 ปี, เกี่ยวข้องในฐานะอื่น), (หญิงอายุตั้งแต่ 15 ปีขึ้นไป เป็นบุตร), (ชายอายุตั้งแต่ 15 ปีขึ้นไป, เป็นบุตร), (ชาย, อายุไม่เกิน 14 ปี, เป็นธิดา) }

A คือ Cartesian Product ของ field ใน record นั่นคือ

$A = A_1 \times A_2 \times A_3$

ให้ A คือเซทของระหัสคำตอบทั้งหมดของ record a ใด ๆ ที่ field i ของ a เป็นอนุเซทของ

สมมุติให้ i = 2 นั่นคือ a มี field ที่ 2 เป็นอนุเซทของ A_2 สมมุติว่าระหัสใน field ที่ 2 ของ a คือ อายุ 15 ปีขึ้นไป นั่นคือ = { 15 + ปี}

ดังนั้นระหัสคำตอบทั้งหมด (all possible code) ของ record a จึงมีอยู่ทั้งสิ้น

$2 \times 1 \times 6 = 12$ ระหัส หรือ

คือ { (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นภรรยา), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็น

สามี), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นบุตร), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นธิดา), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นญาติ), (ชาย, อายุตั้งแต่ 15 ปีขึ้นไป, เกี่ยวข้องในฐานะอื่น ๆ), (หญิง, อายุตั้งแต่ 15 ขึ้นไป, เป็นสามี), (หญิง, อายุตั้งแต่ 15 ปี ขึ้นไป, เป็นภรรยา), (หญิง, อายุตั้งแต่ 15 ปีขึ้นไป, เป็นบุตร), (หญิง, ตั้งแต่ 15 ปีขึ้นไป, เป็นธิดา), (หญิง, อายุตั้งแต่ 15 ปี ขึ้นไป, เป็นญาติ), (หญิง, อายุตั้งแต่ 15 ปีขึ้นไป, เกี่ยวข้องในฐานะอื่น ๆ) }

จึงเห็นได้ว่า ถ้าเรานิยามว่า $A_i^o = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_i^o \times \ldots \times A_n$

ซึ่งในที่นี้คือ $A_2^o = A_1 \times A_2^o \times A_3$ แล้ว A_i^o จะเป็นอนุเซทของ A

คือ $A_i^o \subseteq A$ และขอให้สังเกตไว้เป็น 5 ประการคือ

1. A_2^o คือเซทของระหัสคำตอบทั้งหมดของ record a เมื่อ a คือ record ใด ๆ ที่มี field ที่ 2 ที่เราให้ความสนใจเป็นการเฉพาะเจาะจง

2. ทั้ง A_2^o และ A (ซึ่งกรณีทั่วไปคือ A_i^o และ A) เป็นเซทของเซท หมายความว่าสมาชิกของ A_i^o และ A คือเซทของคำตอบที่พึงเป็นไปได้ของ record a และ record ใด ๆ ตามลำดับ

3. เมื่อสนใจผลลัพธ์หรือคำตอบของเฉพาะ record a จะพบว่าคำตอบของ record a อาจจะเป็นสมาชิกใดสมาชิกหนึ่ง A_i^o (ในที่นี้คือ A_2^o) ก็ได้ แสดงว่า $a \varepsilon a_i^o$

4. A และ A_i^o จะประกอบไปด้วยทั้งเซทของคำตอบที่น่าเป็นไปได้และไม่อาจเป็นไปได้ เช่น (ชาย, อายุไม่เกิน 14 ปี, เป็นภรรยา) เป็นตัวอย่างของคำตอบที่ไม่อาจเป็นไปได้ เรียกว่าเงื่อนไขขัดแย้ง (fail edit)

5. กลุ่มของ Combination ของระหัสคำตอบที่ไม่อาจเป็นไปได้ (Unacceptable) เป็นอนุเซทของ Code Space

ค. สัญญลักษณ์และทฤษฎีที่เกี่ยวข้อง

ค. 1 สัญญลักษณ์

A_i = เซทของคำตอบที่พึงเป็นไปได้ทั้งหมดของรายการข้อมูลที่ i;i = 1, 2, ..., n

n = จำนวนรายการข้อมูลทั้งหมด (จำนวน field)

n_i = จำนวนสมาชิกของเซท A_i; i = 1, 2, ..., n n_i อาจเป็นจำนวนนับได้ (finite) หรือนับไม่ได้ (infinite) ก็ได้

A = Cartesian Product ของทุก field ใน record หรือเซทของระหัสคำตอบที่เป็นไปได้ทั้งหมดของ record

ดังนั้น A = $A_1 \times A_2 \times A_3 \times \ldots \times A_n$

a = record ใด ๆ = เวคเตอร์ที่สมาชิกของ a คือระหัสจากทุก field ของ a[1]

A_i^0 = เซทของระหัสทั้งหมดของ record a ใด ๆ ที่เขาสนใจเฉพาะ field i ของ a ที่เป็นอนุเซทของ A_i หรือ

$A_i^0 = A_1 \times A_2 \times A_3 \times \ldots \times A_{i-1} \times A_i^0 \times \ldots \times A_n$ (; i = 1, 2, ..., n)

A_i^r, A_j^r = field ที่ i ของเงื่อนไขที่ r, field ที่ j ของเงื่อนไขที่ r

ดังนั้น เราสามารถยังผลสรุปได้เป็น 2 ประการดังนี้

1. $a \varepsilon A_i^0$ และระหัสคำตอบของ field ที่ i ของ a จะเป็น Component ของทุกสมาชิกของ A_i^0

2. $A_i^0 \subseteq A_i$ เมื่อนิยามว่า $A_i = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_i \times A_{i+1} \times \ldots \times A_n$

และ $A_i \subseteq A$ เรียก A ว่า Code Space

ค.2 แนวคิดในการสร้างเงื่อนไขการบรรณาธิกรณ์ (Normal form of Edit)

การตรวจสอบหรือบรรณาธิกรณ์ข้อมูลโดยนัยแห่งวิธีการบรรณาธิกรณ์เชิงตรรถวิทยาก็คือ การตรวจสอบโดยสร้างเงื่อนไขการบรรณาธิกรณ์ขึ้นมาใช้ เงื่อนไขจะถูก

---

1. super script "0" หมายความว่า A_i คือ field ที่เรากำลังให้ความสนใจ superscript "i" ชี้ให้เห็นว่าเรากำลังสนใจ field ใดโดยที่ i = 1, 2, ..., n