บทที่ 2

ระเบียบวิธีที่ใช้ในการวิจัย

ในกรณีที่มีค่ารายได้บางตัวขาดหายไปจากการสำรวจตัวอย่าง Greenlees and others (1982: 251-259) ได้เสนอวิธีการประมาณค่ารายได้ที่ขาดหายไปเป็น 2 กรณีใหญ่ ๆ คือ กรณีแรกไม่สนใจกลไก (mechanism) ต่าง ๆ ซึ่งเป็นสาเหตุทำให้ไม่ได้รับคำตอบเกี่ยวกับรายได้ และกรณีที่ 2 เป็นกรณีที่สนใจกลไกอันเป็นสาเหตุทำให้ไม่ได้รับคำตอบเกี่ยวกับรายได้ ดังนี้

2.1 Ignorable Response Mechanism

กรณีนี้เป็นการสมมติว่า การแจกแจงของผู้ตอบ (Response Distribution) ไม่ขึ้นอยู่กับ ค่ารายได้ที่ ได้รับคำตอบ สำหรับวิธีการหาค่ารายได้ที่ขาดหายไปในกรณีนี้ได้อาศัยตัวแปรอื่นของหน่วยตัวอย่างมาช่วยในการวิเคราะห์ ตัวแปรเหล่านี้เรียกว่า auxiliary variables สามารถแบ่งได้เป็น 2 วิธี ดังต่อไปนี้

2.1.1 การแบ่งชั้นภูมิเมื่อเลือกตัวอย่างแล้ว (Post-stratification Approach)

เป็นที่ทราบกันโดยทั่วไปว่า การเลือกตัวอย่างแบบมีชั้นภูมิ (Stratified Sampling) เป็นเทคนิค ที่นิยมใช้อย่างมากในแผนแบบการสำรวจตัวอย่าง เพื่อสนองวัตถุประสงค์ที่ว่า ต้องการให้ตัวอย่างที่ได้มีคุณภาพดีขึ้น และคุณภาพของตัว ประมาณค่าประชากรมีความแม่นยำสูง และการเลือกตัวอย่างแบบมีชั้นภูมินี้จะได้ผลดีที่สุด ถ้าเราสามารถแบ่งประชากรออกเป็นชั้นภูมิโดยคัดหน่วยที่มีลักษณะคล้ายกันเข้าไว้ในชั้นภูมิเดียวกันและให้หน่วยที่อยู่ต่างชั้นภูมิมีลักษณะต่างกัน แต่ในทางปฏิบัติบางครั้งเราไม่สามารถทำได้เช่นนี้ เช่นในกรณีที่เราแบ่งชั้นภูมิออกตามลักษณะภูมิศาสตร์ เช่นแบ่งประเทศไทยออกเป็นภาคต่าง ๆ แต่ละภาคเป็นชั้นภูมิ กรณีนี้เป็นการยากที่ตัวแปรที่เราสนใจจะมีค่าใกล้เคียงกันในแต่ละภาค ประโยชน์ของการเลือกแบบมีชั้นภูมิก็ไม่ใช่มาจากการเพิ่มความแม่นยำของตัว ประมาณ แต่มาจากลักษณะที่สามารถถือภาคแต่ละภาคเป็นอิสระจากกันทำให้ดำเนินการสำรวจได้มีประสิทธิภาพยิ่งขึ้น (สุชาดา กีระนันทน์ 2525: 4.2)

การตอบคำถามของหน่วยตัวอย่าง แต่ละหน่วยอาจขึ้นอยู่กับอายุ เพศ อาชีพ การศึกษาและปัจจัยอื่น ๆ เป็นต้น แต่การใช้ค่าของตัวแปรเหล่านี้เป็นเกณฑ์ในการกำหนดชั้นภูมิก่อนการสุ่มตัวอย่าง อาจทำได้ยาก อย่างไรก็ดี เราอาจใช้ค่าของตัวแปรเหล่านี้ที่ได้จากการสำมะโนครั้งก่อน ๆ มาเป็นเกณฑ์ในการกำหนดชั้นภูมิ หลังจากมีการเลือกหน่วยตัวอย่างด้วยเทคนิคการเลือกตัวอย่าง โดยมีการเก็บข้อมูลที่มีปัจจัยเหล่านี้ในระยะเวลาเดียวกัน (cross-classified) และสามารถทราบขนาดของประชากรในแต่ละชั้นภูมิ (N_h) เพื่อประมาณค่าประชากรรวมของชั้นภูมิต่าง ๆ แล้ว จึงนำมารวมกันเป็นค่าประมาณของประชากร วิธีการเลือกตัวอย่างแบบนี้เรียกว่า "Poststratification" หรือ "stratification after selection" (Holt and Smith 1979: 33)

การแบ่งชั้นภูมิเมื่อเลือกตัวอย่างแล้ว มีประสิทธิภาพดีกว่าการแบ่งชั้นภูมิก่อนเลือกตัวอย่างตรงที่ เราสามารถเลือกปัจจัยที่ใช้ในการแบ่งชั้นภูมิเมื่อเลือกตัวอย่างแล้วคือตัวแปรแบ่งชั้นภูมิ (Stratification Variable) ที่ให้คุณภาพตัวประมาณสูงที่สุด

เทคนิคการแบ่งชั้นภูมิเมื่อเลือกตัวอย่างแล้ว มีประโยชน์ในกรณีที่การสำรวจตัวอย่างมีวัตถุประสงค์ในการสำรวจหลาย ๆ ด้าน (multi-purpose surveys) จนทำให้ตัวแปรแบ่งชั้นภูมิก่อนการเลือกตัวอย่างไม่ค่อยมีความสัมพันธ์กับตัวแปรอื่น ๆ ที่สนใจศึกษา

หลักเกณฑ์ในการเลือกตัวแปรแบ่งชั้นภูมิ เมื่อเลือกตัวอย่างแล้ว Little (1982: 242) ได้กล่าวว่า ให้เลือกตัวแปรที่คิดว่ามีความสัมพันธ์กับความน่าจะเป็นของผู้ตอบและตัวแปรรายได้ที่เราสนใจศึกษา

2.1.1.1 <u>การประมาณค่ารายได้ที่ขาดหายไปโดยใช้ค่าเฉลี่ยจากแต่ละชั้นภูมิ</u>

ค่าเฉลี่ยจากแต่ละชั้นภูมิ คำนวณจากหน่วยตัวอย่างที่ตอบในเรื่องรายได้ของแต่ละชั้นภูมิ

สมมติให้ $\bar{y}_h = \frac{1}{n_h} \sum^{n_h} y_{hi}$

เมื่อ $\bar{y}_h$ เป็นค่าเฉลี่ยรายได้ของหน่วยตัวอย่างที่ตอบในชั้นภูมิ h

y_{hi} เป็นค่ารายได้ของหน่วยที่ i ในชั้นภูมิ h

n_h เป็นจำนวนหน่วยตัวอย่างที่ตอบในชั้นภูมิ h

2.1.1.2 <u>การประมาณค่ารายได้ที่ขาดหายไปโดยใช้ ค่าสุ่มจากแต่ละชั้นภูมิ</u>

วิธีการประมาณค่ารายได้ที่ขาดหายไป โดยใช้ ค่าสุ่มจากแต่ละชั้นภูมิ Greenlees and others (1982: 252) ได้นำวิธีนี้มาจาก Ford (1980) ซึ่งเขาให้ชื่อวิธีนี้ว่า "Hot Deck"

วิธีการแบ่งชั้นภูมิ เมื่อเลือกตัวอย่างแล้ว อาจมีประสิทธิภาพสูงกว่าวิธีการแบ่งชั้นภูมิก่อน เลือกตัวอย่างโดยอาจ เป็นไปได้ในกรณีที่ตัวแปรแบ่งชั้นภูมิก่อนเลือกตัวอย่างไม่ค่อยมีความสัมพันธ์กับกลุ่มตัวแปรที่ต้องการ เก็บข้อมูล แต่วิธีการแบ่งชั้นภูมิ เมื่อ เลือกตัวอย่างแล้วมีโอกาสเลือกตัวแปรแบ่งชั้นภูมิที่เหมาะสมด้วยวิธีต่าง ๆ กัน เพื่อให้ตัวประมาณมีความแม่นยำสูงที่สุด Holt and Smith (1979: 33)

2.1.2 <u>วิธีประมาณค่ารายได้ที่ขาดหายไปจากสมการถดถอยพหุ เชิง เส้น</u>

การใช้ตัวแปรอิสระหลายตัวในการพยากรณ์ตัวแปรตามวิธีหนึ่งที่นิยมใช้ คือ ความถดถอยพหุ เชิง เส้น ซึ่งวิธีการนี้เรียกว่า "Prediction Approach"

สำหรับการประมาณค่าสังเกตของรายได้ที่ขาดหายไป โดยใช้วิธีวิเคราะห์ความถดถอยจะใช้ค่าสังเกตที่มีข้อมูล รายได้ อยู่ทุกหน่วย หาสมการถดถอยโดยวิธีกำลังสองน้อยที่สุด และใช้สมการถดถอยที่ได้ประมาณค่าสังเกตของรายได้ที่ขาดหายไป

กำหนดให้ n = ขนาดตัวอย่างทั้งหมด

n_1 = ขนาดตัวอย่างที่ค่าของ Y ขาดหายไป

$n_r = n - n_1$ = ขนาดตัวอย่างที่มีค่าของ Y และ X ทั้งหมด

จากตัวแบบทั่วไป $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ _____(1)

เมื่อ $\underset{\sim}{Y}$ เป็นเวคเตอร์ของตัวแปรตาม ขนาด $n_r \times 1$

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด $n_r \times p$ และมี full rank $p, p \leq n_r$

$\underset{\sim}{\beta}$ เป็นเวคเตอร์ของพารามิเตอร์ที่ไม่ทราบค่า ขนาด $p \times 1$

$\underset{\sim}{\varepsilon}$ เป็นเวคเตอร์ของความคลาดเคลื่อนขนาด $n_r \times 1$

โดยที่ $E(\varepsilon_i) = 0 \; ; \; i = 1, \ldots, n_r$

$E(\varepsilon_i \varepsilon_j) = \sigma^2 ; \; i = j = 1, \ldots, n_r$

$= 0 \; ; \; i \neq j$

ถ้าประมาณค่าพารามิเตอร์ β ด้วยวิธี กำลังสองน้อยที่สุดจะได้ตัวประมาณ

$$\hat{\beta} = [X'X]^{-1}X' Y$$

และ $\hat{Y} = X\hat{\beta}$ เป็นสมการถดถอยตัวอย่าง _____(2)

โดยที่ $\hat{Y}$ เป็นค่าประมาณของค่าเฉลี่ยของ Y สำหรับค่า X ที่กำหนดให้ (the predicted value of the true mean value of Y for a given X)

หาค่า รายได้ ($\hat{Y}$) ที่ขาดหายไป n_1 หน่วย ได้จาก (2)

สำหรับ วิธี Prediction Approach มีข้อเสียคือ

1) ถ้า ความน่าจะเป็นของการตอบเรื่องรายได้ขึ้นอยู่กับ ระดับรายได้แล้ว ตัวประมาณ β ที่ได้ จะผิดพลาด เนื่องจากค่ารายได้ เป็นตัวแปรตามในสมการถดถอย

2) $\hat{Y}_i$ มีค่าไม่เท่ากับ $X_i \hat{\beta}$ เมื่อ $i = n_{r+1}, \ldots, n$ ถึงแม้ว่า $\hat{\beta}$ จะเป็นตัวประมาณที่ไม่เอนเอียง (unbiased estimator) เนื่องจากเราไม่ได้นำค่าของตัวอย่างเหล่านี้ไปใช้ในการหาสมการถดถอย

2.2 Nonignorable Response Mechanism

กรณีนี้เป็นการสมมติว่า ความน่าจะเป็นในการตอบเรื่อง รายได้ ขึ้นอยู่กับระดับรายได้ ของผู้ตอบ Greenlees and others (1982: 252-254) ได้เสนอวิธีการดังต่อไปนี้

(1) การประมาณค่าพารามิเตอร์ β ที่ได้จากสมการถดถอยด้วยวิธีแมกซิมัมไลลิฮูด

กำหนดให้ Y แทนตัวแปรรายได้

Z แทนตัวแปรอื่น ๆ (other variables)

สมมติให้ ความน่าจะเป็นของการตอบคำถามเกี่ยวกับ Y ขึ้นอยู่กับค่า Y โดยที่มีลักษณะเป็น logistic function ของ Y และตัวแปร Z ดังนี้

$$P(R_i = 1 \mid Y_i, Z_i) = \frac{1}{1+\exp(-\alpha-\gamma Y_i - Z_i\delta)} \qquad (3)$$

โดยที่ R_i = 1 ถ้า i เป็นหน่วยที่ตอบค่า Y

= 0 ถ้า i เป็นหน่วยที่ไม่ตอบค่า Y

Z_i เป็นเวคเตอร์ของหน่วยที่ i ขนาด $1 \times n_r$

α, γ เป็นพารามิเตอร์ที่ไม่ทราบค่า (scalar parameters)

δ เป็น เวคเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $n_r \times 1$

และจากตัวแบบทั่วไป ของสมการถดถอย

$$Y_i = X_i\beta + \varepsilon_i \qquad (4)$$

ถ้า (3) และ (4) เป็นจริงสำหรับตัวอย่างสุ่มขนาด n

เมื่อ n เป็นจำนวนตัวอย่างทั้งหมด

n_r เป็นจำนวนตัวอย่างที่ตอบคำ Y_i

และ X_i เป็นเวคเตอร์ของตัวแปรอิสระขนาด $1 \times p$

$i = 1, 2, \cdots, n_r$ เป็นลำดับของ ตัวอย่างที่ตอบคำ Y_i

$i = n_{r+1}, \cdots, n$ เป็นลำดับของ ตัวอย่างที่ไม่ตอบคำ Y_i

ฟังก์ชั่นภาวะน่าจะเป็นสำหรับตัวอย่างสุ่มขนาด n คือผลคูณของ $(n-n_r)$ factors และ n_r factors

โดยที่ $$L_i = \frac{1}{1+\exp(-\alpha-\gamma Y_i - Z_i\delta)} \times \frac{1}{\sigma}\ \phi\left(\frac{Y_i - X_i\beta}{\sigma}\right) \qquad (5)$$

$$, \quad i = 1, \cdots, n_r$$

เป็น ภาวะน่าจะเป็น (likelihood) ของหน่วย i ที่ตอบ คำ Y

และ $$L_i = \int_{-\infty}^{\infty} \left(1 - \frac{1}{1+\exp(-\alpha-\gamma Y - Z_i\delta)}\right) \qquad (6)$$

$$\times \frac{1}{\sigma}\ \phi\left(\frac{Y - X_i\beta}{\sigma}\right) dY \quad , \qquad i = n_{r+1}, \cdots, n$$

เป็น ภาวะน่าจะเป็น(likelihood)ของหน่วย i ที่ไม่ตอบคำ Y

∴ ฟังก์ชั่นภาวะน่าจะเป็นของตัวอย่างสุ่มขนาด n คือ $L = \prod_{i=1}^{n} L_i$ เมื่อ L_i เป็นค่าที่กำหนดโดย (5) หรือ (6) ตามค่าของ i

ด้วยวิธีแมกซิมัมไลลิฮูด จะได้ $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{\delta}$ และ $\hat{\sigma}$ เป็นตัวประมาณค่าของ α, β, γ, δ และ σ ตามลำดับ

(2) การประมาณค่ารายได้ที่ไม่ได้รับคำตอบ

กำหนดให้ $E(Y_i | X_i, Z_i, R_i = 0) \quad i = n_{r+1}, ---, n$

เป็นค่าประมาณของ Y_i ที่ไม่ได้รับคำตอบ สูตรในการคำนวณเป็นดังนี้

$$E(Y_i | X_i, Z_i, R_i = 0) = \frac{-\int_{-\infty}^{\infty} Y\left(1 - \frac{1}{1+\exp(-\hat{\alpha}-\hat{\gamma}Y-Z_i\hat{\delta})}\right)}{-\int_{-\infty}^{\infty} \left(1 - \frac{1}{1+\exp(-\hat{\alpha}-\hat{\gamma}Y-Z_i\hat{\delta})}\right)}$$

$$\times \frac{\frac{1}{\hat{\sigma}} \phi\left(\frac{Y-X_i\hat{\beta}}{\hat{\sigma}}\right) dY}{\frac{1}{\hat{\sigma}} \phi\left(\frac{Y-X_i\hat{\beta}}{\hat{\sigma}}\right) dY} \qquad \text{----------(7)}$$

ข้อเสียของการใช้ $E(Y_i | X_i, Z_i, R_i=0)$ ประมาณค่า Y ที่ขาดหายไปคือต้องคำนึงถึงค่า V(Y) ว่าให้ค่าความแปรปรวนต่ำกว่าวิธีอื่นหรือไม่ เพื่อหลีกเลี่ยงการกล่าวถึงค่า V(Y) จะใช้วิธีทำซ้ำ ๆ (iterative) เพื่อประมาณค่า Y ดังต่อไปนี้

1. สุ่ม ε_i จาก N(0,1) generator

2. คำนวณ $Y_i = X_i\hat{\beta} + \hat{\sigma}\varepsilon_i$ และความน่าจะเป็นของการไม่ตอบ

$$P(R=0 | Y_i, Z_i) = 1-1/\left[1+\exp(-\hat{\alpha}-\hat{\gamma}Y_i-Z_i\hat{\delta})\right]$$

3. สุ่มตัวแปรสุ่ม (random variable) η จาก uniform generator ในช่วง $[0,1]$

4. Y_i จะเป็นตัวประมาณที่ขาดหายไป (imputed value) สำหรับค่าสังเกตที่ i ถ้า $P(R=0 | Y_i > Z_i) \geq \eta$

5. ทำขั้นตอนที่หนึ่งซ้ำ

2.3 การเลือกวิธีที่ใช้ในการหาค่ารายได้ที่ขาดหายไป

ในการวิจัยครั้งนี้ ได้เลือกใช้วิธีประมาณค่ารายได้ที่ขาดหายไปในกรณีที่ไม่สนใจกลไกที่เป็นสาเหตุของการตอบเรื่องรายได้ (Ignorable Response Mechanism) ดังกล่าวมาแล้วข้างต้น ด้วยเหตุผลดังต่อไปนี้

1. ปัญหาทางด้านข้อมูล

เนื่องจากในประเทศไทย ยังไม่มีบริการด้านเผยแพร่ข้อมูลให้แก่นักวิจัยดังเช่นในบางประเทศ และการวิจัยครั้งนี้เป็นการวิจัยเพื่อหาวิธีที่เหมาะสมกับลักษณะของข้อมูลภายในประเทศ ดังนั้น ลักษณะบางอย่างของข้อมูลที่นักสถิติชาวต่างประเทศใช้วิจัยจึงไม่อาจนำมาประยุกต์ใช้ได้ ในประเทศสหรัฐอเมริกา นักสถิติสามารถขอข้อมูลเกี่ยวกับค่าจ้างและเงินเดือนจาก IRS (Internal Revenue Service) เพื่อนำมาศึกษาเกี่ยวกับปัญหารายได้ที่ขาดหายไป กล่าวคือสามารถนำหน่วยตัวอย่างที่ได้จากการสำรวจ CPS (Current Population Survey) มาจับคู่กับข้อมูลจาก IRS เพื่อหาหน่วยตัวอย่างที่ไม่ตอบในเรื่องรายได้ ตามที่มีอยู่จริง ๆ เพื่อนำข้อมูลจากหน่วยตัวอย่างทั้งที่ให้คำตอบและไม่ให้คำตอบในเรื่องรายได้ มาทดสอบดูว่าความน่าจะเป็นของการตอบขึ้นอยู่กับรายได้หรือไม่

2. ด้วยเหตุผลที่ว่า เราไม่สามารถหาข้อมูลรายได้ที่ไม่ได้รับคำตอบจากหน่วยงานที่เกี่ยวข้องกับการเก็บรวบรวมข้อมูลในด้านนี้ในทางปฏิบัติจึงไม่สามารถหาได้ว่าความน่าจะเป็นของการตอบขึ้นอยู่กับรายได้หรือไม่ และถ้ามันขึ้นอยู่กับรายได้จริง ๆ แล้ว รูปแบบของความสัมพันธ์จะเป็นไปในลักษณะใด จึงไม่อาจพิจารณากรณีที่สนใจกลไกที่เป็นสาเหตุของการตอบเรื่องรายได้ (Nonignorable Response Mechanism)

3. ปัญหาด้านเอกสารที่ใช้ประกอบการวิจัย

สำหรับเอกสารที่ใช้ประกอบการค้นคว้าวิจัยเกี่ยวกับเรื่องรายได้ที่ขาดหายไปในการวิจัยครั้งนี้เนื่องจากเอกสารส่วนใหญ่เป็นการสัมมนาทางวิชาการในต่างประเทศ ซึ่งยังไม่เผยแพร่ในประเทศไทย

4. วิธีประมาณค่ารายได้ที่ขาดหายไปในกรณีที่ไม่สนใจกลไกที่เป็นสาเหตุของการตอบเรื่องรายได้ เป็นวิธีที่คำนวณได้ง่าย และเป็นวิธีที่นิยมใช้ในการประมาณค่าตัวแปรที่ขาดหายไป

ในการหาวิธีที่เหมาะสมที่สุดสำหรับการหาค่ารายได้ที่ขาดหายไปจากวิธีต่าง ๆ ที่นำมาศึกษาวิจัยครั้งนี้ จะพิจารณาจากร้อยละของความแตกต่างระหว่างค่าจริงกับค่าประมาณของรายได้ และจากการทดสอบสมมติฐาน เพื่อทดสอบความแตกต่างระหว่างค่าจริงกับค่าประมาณของรายได้ที่ได้จากแต่ละวิธีและทดสอบความแตกต่างระหว่างค่าประมาณรายได้ที่ได้จากแต่ละวิธี โดยการทดสอบแบบจับคู่สิ่งทดลอง