บทที่ 3
ระเบียบวิธีวิจัย

3.1 วิธีการเลือกตัวอย่าง

วิธีการเลือกตัวอย่างสำหรับการศึกษานี้ใช้วิธีการสุ่มตัวอย่างแบบง่าย (Simple Random Technique) ดังนี้คือ

3.1.1 กรณีผู้ประกอบการที่มีสถานประกอบการตั้งอยู่ในเขตอำเภอเมือง จังหวัดเชียงใหม่ ได้เก็บข้อมูลจากสถิติการจดทะเบียนจัดตั้งห้างหุ้นส่วนจำกัดและบริษัทจำกัดจากพาณิชย์จังหวัดเชียงใหม่โดยมีจำนวนตัวอย่างที่เลือกดังนี้

ห้างหุ้นส่วนจำกัด	55	ราย
บริษัทจำกัด ที่มีผู้ถือหุ้นเป็นคนไทยล้วน	40	
บริษัทจำกัด ที่มีผู้ถือหุ้นเป็นคนไทยและคนต่างชาติ	15	
รวม	110	ราย

3.1.2 กรณีผู้ประกอบการที่มีสถานประกอบการตั้งอยู่ในนิคมอุตสาหกรรมภาคเหนือ จังหวัดลำพูนได้เก็บข้อมูลจากรายละเอียดเกี่ยวกับผู้ซื้อที่ดินในนิคมอุตสาหกรรมภาคเหนือโดยเลือกเฉพาะผู้ซื้อที่ดินที่เปิดดำเนินการแล้วเท่านั้น และเนื่องจากกิจการที่เปิดดำเนินการแล้วเป็นบริษัทจำกัดทั้งสิ้นรวม 60 บริษัทโดยไม่มีห้างหุ้นส่วนจำกัดเปิดดำเนินการเลย และใน 60 บริษัท ดังกล่าวประกอบด้วย

บริษัทจำกัดที่มีผู้ถือหุ้นเป็นคนไทยล้วน	17	ราย
บริษัทจำกัดที่มีผู้ถือหุ้นเป็นคนไทยและคนต่างชาติ	43	
รวมจำนวนผู้ประกอบการที่มี	60	ราย

ดังนั้นในการเลือกตัวอย่างได้ดำเนินการเลือกดังนี้

บริษัทจำกัดที่มีผู้ถือหุ้นเป็นคนไทยล้วน	15	ราย
บริษัทจำกัดที่มีผู้ถือหุ้นเป็นคนไทยและคนต่างชาติ	40	
รวมจำนวนตัวอย่างที่เลือก	55	ราย

3.2 ทฤษฎีการประมาณค่าที่ใช้ในการศึกษา

เนื่องจากงานบริการทางด้านบัญชีเป็นงานบริการที่มีผลกระทบต่อการดำรงอยู่และความเจริญเติบโตของกิจการโดยตรง ซึ่งค่าธรรมเนียมบริการของบริษัทที่ปรึกษาทางด้านบัญชีแต่ละแห่งนั้นมีวิธีการคิดที่แตกต่างกัน ยกตัวอย่างเช่น ถ้าเป็นบริษัทที่ปรึกษาทางด้านบัญชีที่ได้มาตรฐานอยู่ในระดับนานาชาติ ซึ่งปัจจุบันมีอยู่เพียง 5 แห่งดังที่กล่าวมาแล้วข้างต้น ก็จะมีการคิดค่าบริการในอัตราที่สูงมาก ดังนั้นในการตัดสินใจเลือกใช้บริการที่ปรึกษาทางด้านบัญชี ผู้ประกอบการจำเป็นต้องคำนึงถึงประโยชน์ที่ตนเองจะได้รับก่อนการตัดสินใจเลือกใช้บริการระหว่างบริษัทที่ปรึกษาทางด้านบัญชีที่ได้มาตรฐานเหล่านั้นกับบริษัทที่ปรึกษาทางด้านบัญชีอื่น ๆ อย่างไรก็ตาม ปัจจุบันนี้บริษัทที่ปรึกษาทางด้านบัญชีมีจำนวนอยู่มากมายทั้งในจังหวัดเชียงใหม่เองและในจังหวัดอื่นๆซึ่งสามารถให้บริการแก่ผู้ประกอบการที่ตั้งอยู่ในจังหวัดเชียงใหม่ และในนิคมอุตสาหกรรมภาคเหนือได้ทั้งสิ้น ดังนั้นผู้ประกอบการจึงมีทางเลือกหลากหลายในการเลือกใช้บริการจากสำนักงานหรือบริษัทที่ปรึกษาทางด้านบัญชีดังกล่าว ในการศึกษาครั้งนี้จึงได้กำหนดกรอบของการศึกษาในส่วนของปัจจัยหรือตัวแปรที่มีอิทธิพลต่อการตัดสินใจเลือกใช้บริการจากบริษัทที่ปรึกษาทางด้านบัญชี ทั้งที่จัดอยู่ในกลุ่มที่ได้มาตรฐาน และที่ไม่ได้มาตรฐานซึ่งมีอยู่จำนวนมาก และเนื่องจากลักษณะของการศึกษานี้ ตัวแปรตามจะเป็นตัวแปรหุ่น (dummy variable) ทรงศักดิ์ ศรีบุญจิตต์ (2543) ได้กล่าวไว้ดังนี้

"ในการศึกษาจำนวนไม่น้อย เราจะพบว่าตัวแปรตาม (dependent variable) จะมีลักษณะเป็นทางเลือกเชิงคุณภาพ (qualitative choice) เป็น 2 ทางเลือกหรือมากกว่า เช่น การเลือกตั้ง การยอมรับเทคโนโลยีของเกษตรกร การเข้าเป็นสมาชิกสหกรณ์การเกษตรของเกษตรกร การเข้าเป็นสามชิกกลุ่มแม่บ้านของแม่บ้านเกษตรกร การเลือกวิธีเดินทางไปทำงานว่าเป็นทางรถเมล์ รถไฟ รถยนต์ หรือจักรยาน เป็นต้น คำถามที่เกิดขึ้นก็คือว่าแบบจำลองที่มีตัวแปรตามเป็นลักษณะเช่นนี้ เราจะทำการประมาณค่าได้หรือไม่ หรือว่าเราต้องใช้วิธีการอื่นเป็นการเฉพาะ รวมทั้งการลงความเห็นในเชิงสถิติด้วยว่ามีปัญหาหรือไม่และจะต้องแก้ไขอย่างไร แบบจำลอง 4 แบบ ซึ่งคือ (1) แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) (2) แบบจำลองโพรบิท (probit model) (3) แบบจำลองโลจิท (logit model) และ (4) แบบจำลองโทบิท (tobit model) ซึ่งจะเสนอนี้จะตอบคำถามดังกล่าวข้างบน

สำหรับเนื้อหาของแบบจำลองเหล่านี้มีรายละเอียดดังนี้คือ

แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model)

สมมติว่าเรามีแบบจำลองอย่างง่ายดังนี้

$$y_i = \alpha + \beta x_i + u_i \qquad (1)$$

โดยที่ y_i = 1 ถ้าครัวเรือนที่ i ซื้อรถยนต์ (ซึ่งอาจเป็นตัวแปรตามในลักษณะอื่น ๆ อีกก็ได้ เช่นถ้าครัวเรือนที่ i ซื้อบ้าน หรือครัวเรือนเกษตรกร ครัวเรือนที่ i ได้รับเอาเทคโนโลยีชนิด ก. มาใช้ในการผลิตเป็นต้น)

y_i = 0 ถ้าครัวเรือนที่ i ไม่ซื้อรถยนต์ (หรือครัวเรือนที่ i ไม่ซื้อบ้าน หรือเกษตรกรครัวเรือนที่ i ไม่รับเอาเทคโนโลยีชนิด ก. มาใช้ในการผลิต ตามตัวอย่างข้างบน)

u_i = ตัวแปรสุ่ม (random variable) หรือเทอมคลาดเคลื่อน (error terms) หรือตัวรบกวน (disturbances) ที่มีการแจกแจงเป็นอิสระและมีค่าเฉลี่ยเท่ากับศูนย์

แบบจำลองตามสมการ (1) นี้เรียกว่า "แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) สำหรับคำอธิบายว่าทำไมเราจึงเรียกว่าเป็นแบบจำลองความน่าจะเป็นเชิงเส้นนั้น จะได้กล่าวถึงในลำดับต่อไป

จากสมการเราสามารถหาค่าคาดหมายแบบมีเงื่อนไข (conditional expected value) ของค่าสังเกตของตัวแปรตามแต่ละตัว y_i โดยกำหนดค่าตัวแปรอธิบาย (Explanatory variable) หรือตัวแปรอิสระ (independent variable) ในกรณีนี้ซึ่งคือ x_i มาให้ได้ดังนี้

$$E(y_i / x_i) = \alpha + \beta x_i \qquad (2)$$

และเนื่องจาก y_i มีค่าเพียง 2 ค่าเท่านั้นดังได้กล่าวไว้ข้างต้นคือ 1 และ 0 เพราะฉะนั้นเราสามารถที่จะหาการแจกแจงความน่าจะเป็นของ y_i ได้โดยการให้

p_i = ความน่าจะเป็นที่ $y_i = 1$ ซึ่งเขียนแทนด้วยสัญลักษณ์ $p_i = \mathbf{prob}(y_i = 1)$

และ

$1 - p_i$ = ความน่าจะเป็นที่ $y_i = 0$ ซึ่งเขียนแทนด้วยสัญลักษณ์ $p_i = \mathbf{prob}(y_i = 0)$

ซึ่ง y_i ก็จะมีการแจกแจงความน่าจะเป็น (probability distribution) ดังนี้

y_i	ความน่าจะเป็น (probability)
0	$1-p_i$
1	p_i

จากการแจกแจงความน่าจะเป็นดังกล่าว เราสามารถหาค่าคาดหมาย (expected value) ของ y_i ได้ดังนี้

$$E(y_i) = 1(p_i) + 0(1-p_i) = p_i \tag{3}$$

จะเห็นได้ว่าค่าคาดหมาย (expected value) ของ y_i จากสมการ (2) และ (3) คือค่าเดียวกันเพราะฉะนั้นสมการ (2) และ (3)จึงเท่ากันเพราะฉะนั้นเราจะได้

$$p_i = \alpha + \beta x_i = E(y_i / x_i) \tag{4}$$

นั่นคือค่าคาดหมายแบบมีเงื่อนไข (conditional expectation) ของ y_i จากแบบจำลอง (1) คือ ความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ของ y_i นั่นเอง (Gujarati, 1995, PP 540 – 542; Pindyck และ Rubinfeld, 1998, PP 298 – 300) เพราะฉะนั้นจึงเป็นการตอบคำถามว่าเราจึงเรียกแบบจำลอง (1) ว่าเป็นแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability) เพราะฉะนั้นโดยสรุปแล้วเรามักจะเขียนแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) โดยให้ตัวแปรตามเป็นความน่าจะเป็น (probability) ได้ดังนี้

$$p_i = \begin{cases} \alpha + \beta x_i & \text{เมื่อ } 0 < \alpha + \beta x_i < 1 \\ 1 & \text{เมื่อ } \alpha + \beta x_i > 1 \\ 0 & \text{เมื่อ } \alpha + \beta x_i < 0 \end{cases}$$

(Pindyck และ Rubinfield, 1998, P300)

ปัญหาในการประมาณค่าแบบจำลองความน่าจะเป็น (linear probability model)

คำถามที่ตามมาก็คือว่าเราจะใช้วิธีการกำลังสองน้อยที่สุดแบบธรรมดา (Ordinary Least Squares, OLS) ได้หรือไม่ (Gujarati, 1995, PP 542 – 546; Pindych และ Rubinfeld, 1998, PP 300 - 304; Johnston และ Dinardo, 1997, PP 414 – 418) ได้กล่าวถึงปัญหาในการประมาณค่าแบบจำลองความน่าจะเป็นเชิงเส้นดังนี้

1. ปัญหาการแจกแจงแบบไม่ปกติ (non-normality)

โดยทฤษฎีแล้วเราทราบว่าตัวประมาณค่า OLS (OLS estimator) นั้นหามาได้โดยไม่ต้องใช้ข้อสมมติเกี่ยวกับการแจกแจงแบบปกติของ u_i แต่อย่างไรก็ตามในการลงความเห็นในเชิงสถิติ (statistical inference) เป็นต้นว่า การทดสอบสมมุติฐาน ฯลฯ เราก็ยังคงต้องใช้ข้อสมมติ (assumption) เกี่ยวกับการแจกแจงปกติของ u_i อยู่ดี แต่ข้อสมมติเกี่ยวกับการแจกแจงปกติของ u_i นี้ไม่เป็นจริง ในกรณีของแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) เพราะว่า u_i (ซึ่งเหมือนกับ y_i) จะมี 2 ค่าเท่านั้น โดยพิจารณาจาก

$$u_i = y_i - \alpha - \beta x_i \tag{5}$$

ซึ่งจะเห็นได้ว่าเมื่อ $y_i = 1$ จะได้ $u_i = 1 - \alpha - \beta x_i$ (6)

และเมื่อ $y_i = 0$ จะได้ $u_i = - \alpha - \beta x_i$ (7)

ซึ่งจะเห็นได้ว่า u_i จะไม่มีการแจกแจงแบบปกติ ซึ่งแท้ที่จริง u_i มีการแจกแจงแบบ binomial distribution (Gujarati, 1995, PP 542 – 543) อย่างไรก็ตามการที่ข้อสมมุติเกี่ยวกับการแจกแจงปกติของ u_i ไม่เป็นจริงดังที่ปรากฏนั้นอาจจะไม่ใช่สิ่งที่สำคัญนัก เพราะว่าเราทราบว่าค่าประมาณแบบจุดด้วยวิธี OLS (OLS point estimates) ยังคง "ไม่เอนเอียง (unbiased)" ประกอบกับเมื่อขนาดของตัวอย่างเพิ่มขึ้นอย่างไม่จำกัดเราสามารถจะพิสูจน์ได้ว่าตัวประมาณค่า OLS มีแนวโน้มที่จะมีการแจกแจงปกติ เพราะฉะนั้นในกรณีที่ตัวอย่างมีขนาดใหญ่การลงความเห็นทางสถิติ (statistical inference) เกี่ยวกับแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) ก็จะเป็นตามกระบวนการของ OLS ภายใต้ข้อสมมุติเกี่ยวกับการแจกแจงปกติของ u_i (Gujarati, 1995, P 543) แต่ก็จะต้องมีการแก้ heteroscedasticity ในกระบวนการประมาณค่าซึ่งก็คือการใช้แนวทาง GLS (Generalized Least Squares) หรือวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weight least squares) เนื่องด้วยเหตุผลดังต่อไปนี้

2. ความแปรปรวนของเทอมความคลาดเคลื่อน (error terms) มีลักษณะ (heteroscedastic)

การแจกแจงความน่าจะเป็นของเทอมความคลาดเคลื่อน (error terms) u_i สามารถหาได้โดยการแทนค่า y_i ด้วย 1 และ 0 ตามลำดับเข้าไปในแบบจำลอง (1) จะได้

$1 = \alpha + \beta x_i + u_i$ ซึ่งคือ $u_i = 1 - \alpha - \beta x_i$ (8)

$0 = \alpha + \beta x_i + u_i$ ซึ่งคือ $u_i = - \alpha - \beta x_i$ (9)

เพราะฉะนั้นการแจกแจงความน่าจะเป็นของ u_i สามารถเขียนได้ดังนี้

y_i	u_i	ความน่าจะเป็น
1	$1 - \alpha - \beta x_i$	p_i
0	$- \alpha - \beta x_i$	$1 - p_i$

และจากข้อสมมุติที่ว่าเทอมความคลาดเคลื่อน (error terms) มีค่าเฉลี่ยเป็นศูนย์เราจะได้ว่า

$$E(u_i) = (1 - \alpha - \beta x_i)p_i + (- \alpha - \beta x_i)(1 - p_i) = 0 \tag{10}$$

และหาค่าของ p_i และ $1 - p_i$ จะได้

$$p_i = \alpha + \beta x_i \tag{11}$$

$$1 - p_i = 1 - \alpha - \beta x_i \tag{12}$$

และจากคำนิยามของความแปรปรวนที่ใช้กับเทอมความคลาดเคลื่อนที่มีค่าเฉลี่ยเท่ากับศูนย์จะได้ว่า

$$\begin{aligned} Eu_i^2 &= (1 - \alpha - \beta x_i)^2 p_i + (-\alpha - \beta x_i)^2 (1 - p_i) \\ &= (1 - \alpha - \beta x_i)^2 (\alpha + \beta x_i) + (\alpha + \beta x_i)^2 (1 - \alpha - \beta x_i) \\ &= (1 - \alpha - \beta x_i)(\alpha + \beta x_i) = p_i(1 - p_i) \end{aligned} \tag{13}$$

ซึ่งก็คือ $$Eu_i^2 = \sigma_i^2 = \mathrm{var}(u_i) = E(y_i / x_i)[1 - E(y_i / x_i)] = p_i(1 - p_i) \tag{14}$$

(Gujarati, 1995, P 543; Pindyck และ Rubinfeld, 1998, P 300)

สมการ (14) แสดงให้เห็นว่าเทอมความคลาดเคลื่อน (error term) มีลักษณะที่ว่าความแปรปรวนของเทอมความคลาดเคลื่อน (error term) แตกต่างกัน (heteroscedastic) ค่าสังเกตที่มีค่า p_i เข้าใกล้ 0 หรือ 1 จะมีค่าความแปรปรวนโดยเปรียบเทียบต่ำในขณะที่ค่าสังเกตที่มี p_i ใกล้ 0.5 จะมีความแปรปรวนสูงกว่า (Pindyck และ Rubinfeld, 1998, P 300) และการมีความแปรปรวนแตกต่างกัน (heteroscedasticity) จะทำให้ตัวประมาณค่า OLS (OLS estimator) มีประสิทธิภาพต่ำกว่าตัวประมาณค่า GLS (GLS estimator) แต่ทว่าตัวประมาณค่า OLS (OLS estimator) ก็ยังคง "คล้องจอง (consistent)" และ "ไม่เอนเอียง (unbiases)" อยู่ (Dhryme; Rubinfeld, 1998, P 300) และวิธีแก้ความแปรปรวนไม่เท่ากัน (heteroscedasticity) ก็คือการประมาณค่าความแปรปรวน (variance) ของแต่ละค่าของ y_i และใช้การประมาณค่าแบบกำลังสองน้อยที่สุดถ่วงน้ำหนัก (weighted least squares estimation) ซึ่งสามารถทำได้โดยใช้วิธีกำลังสองน้อยที่สุดธรรมดา (ordinary least squares)

ประมาณค่าแบบจำลอง (1) ซึ่งจะทำให้เราได้ $\hat{y}_i$ ซึ่งก็คือค่าประมาณของ $E(y_i/x_i)$ ซึ่งแท้จริง (true $E(y_i/x_i)$) และจากสมการ (14) เราสามารถประมาณค่าความแปรปรวนของความคลาดเคลื่อน (error) แต่ละตัวดังนี้

$$\hat{\sigma}_i^2 = \hat{y}_i(1-\hat{y}_i) \quad \text{โดยที่} \quad \hat{y}_i = \hat{\alpha}+\hat{\beta}x_i \tag{15}$$

(Pindyck และ Rubinfeld, 1998, PP 300 –301; Gujarati, 1995, P 544)

นั่นคือเราต้องแปลง (transform) ข้อมูลโดยการหารทั้งสองข้างของแบบจำลอง (1) ของเราด้วย $\hat{\sigma}i$ ซึ่งจะได้ดังนี้

$$\frac{y_i}{\hat{\sigma}_i} = \frac{\alpha}{\hat{\sigma}_i} + \frac{\beta x_i}{\hat{\sigma}_i} + \frac{u_i}{\hat{\sigma}} \tag{16}$$

ซึ่งจะทำให้ค่าความคลาดเคลื่อนของแบบจำลอง (16) มีความแปรปรวนเท่ากันหรือเหมือนกัน (homoscedastic) ซึ่งเราก็สามารถใช้วิธี OLS เข้ากับแบบจำลอง (16) ได้เลย (Gujarati, 1995, P 544; Johnston, 1984, P 302) อย่างไรก็ตาม Pindyck และ Rubinfeld (1998, P 301) กล่าวว่าสิ่งที่เป็นปัญหาเกี่ยวกับวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weighted least squares) ก็คือว่าไม่มีการรับประกันได้ว่าค่าทำนาย (predicted value) $\hat{y}_i$ จะอยู่ในช่วง (0,1) ถ้าค่าของ $\hat{y}_i$ อยู่นอกช่วง (0,1) ค่าสังเกต (observation) นั้นก็จะถูกเอาออกจากแบบจำลองหรือไม่ก็อาจจะต้องให้มีค่าดังเช่น .01 และ .99 และไม่ว่าจะเป็นกรณีใดก็ตามกระบวนการกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weighted least squares) จะไม่มีประสิทธิภาพสำหรับ finite samples และเนื่องจากกระบวนการกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักมีความอ่อนไหวต่อความคลาดเคลื่อนของ แบบจำลองที่เราระบุขึ้นมา (errors of specification) Pindyck และ Rubinfeld (1998, P 301) จึงแนะนำไม่ให้ใช้กระบวนการวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weighted least squares procedure)

3. ปัญหา $\hat{y}_i$ ออกนอกช่วง 0 และ 1 ซึ่งไม่สอดคล้อง $0 \leq E(y_i/x_i) \leq 1$

Johnston และ Dinardo (1997, P 417) และ Pindyck และ Rubinfeld (1998, P 301) กล่าวว่า จุดอ่อนที่สำคัญมากของแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) ก็คือว่าแบบจำลองนี้ไม่ได้บังคับ (constrain) ให้ค่าทำนาย (ซึ่งคือ $\hat{y}_i$) ตกอยู่ในช่วง 0 และ 1 ทั้ง ๆ ที่โดยทฤษฎีแล้ว $E(y_i/x_i)$ ในแบบจำลองความน่าจะเป็นเชิงเส้นซึ่งวัดความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ (event) y ที่เกิดขึ้นเมื่อ x ถูกกำหนดมาให้จะต้องตกอยู่ระหว่าง 0 และ 1 แต่ก็ไม่มี

อะไรมารับประกันได้ว่า $\hat{y}_i$ [ซึ่งก็คือตัวประมาณค่า (estimators) ของ $E(y_i/x_i)$] จะอยู่ในช่วง 0 และ 1 ดังกล่าว นี่คือปัญหาที่แท้จริงที่ใช้วิธี OLS ในการประมาณค่าแบบจำลองความน่าจะเป็นเชิงเส้น (Gujarati, 1995, P 544) วิธีแก้ปัญหานี้ก็คือถ้าค่า $\hat{y}_i$ ใดมีค่าน้อยกว่า 0 ก็ให้ถือว่า $\hat{y}_i$ นั้นมีค่าเท่ากับ 0 (Gujarati, 1995 P 544; Pindyck และ Rubinfeld, 1998, P 301) แต่วิธีนี้ Pindyck และ Rubinfeld (1998, P 301) กล่าวว่าไม่ใช่วิธีที่เป็นที่น่าพอใจมาก เพราะว่าเราอาจจะทำนายเหตุการณ์ว่าจะเกิดขึ้นด้วยความน่าจะเป็นเท่ากับ 0 ในขณะที่ความจริงแล้วเหตุการณ์นั้นอาจจะเกิดขึ้นก็ได้ และข้อสังเกตอีกประการหนึ่งที่ Rubinfeld (1998, P 301) ได้ให้ไว้ก็คือในขณะที่กระบวนการประมาณค่าอาจจะให้ค่าประมาณที่ "ไม่เอนเอียง (unbiased)" ค่าทำนาย (prediction) ที่ได้รับจากกรรมวิธีประมาณค่าจะมีลักษณะ "เอนเอียง (biased)" อย่างชัดเจน

Pindyck และ Rubinfeld, (1998, P 301) ให้ข้อเสนอแนะเพิ่มเติมว่าวิธีการอีกวิธีหนึ่งที่จะแก้ปัญหาการที่ค่า $\hat{y}_i$ มีค่าออกนอกช่วง 0 และ 1 ก็คือการประมาณค่าพารามิเตอร์ α และ β ใหม่ โดยมีข้อจำกัดหรือข้อบังคับ (constraint) ว่า $0 \leq \hat{y}_i \leq 1$ แต่อย่างไรก็ตาม เนื่องจากไม่มีอะไรรับประกันว่าค่าประมาณ (estimates) จะมีลักษณะ "ไม่เอนเอียง (unbiased)" จึงคิดว่าการใช้วิธีการกำลังสองน้อยที่สุดแบบธรรมดา (ordinary least squares) กับแบบจำลองความน่าจะเป็นเชิงเส้นน่าจะเหมาะสมกว่า

4. ปัญหาการประมาณค่าสโลป (slope) ที่สูงเกินจริง (overestimated slope) หรือต่ำเกินจริง (underestimated slope)

ปัญหาที่สำคัญมากอีกปัญหาหนึ่งของการประมาณค่า (estimation) แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) ด้วยวิธีกำลังสองน้อยที่สุดธรรมดา (ordinary least squares) ก็คือค่าของสโลปที่ประมาณค่าได้อาจจะมีค่าสูงเกินความเป็นจริง (overestimated slope) หรือต่ำกว่าความเป็นจริง (underestimated slope) ได้ ถ้าหากว่าค่าสังเกต (observations) ที่เลือกมาหรือได้มานั้นมีคุณลักษณะประจำตัว (คือค่า x) ที่มีค่าสุดโต่งหรือปลายสุด (extreme values) เป็นจำนวนมากเกินไปทำให้ได้ค่าประมาณของสโลป (slope estimated) จากวิธีกำลังสองน้อยที่สุดแบบธรรมดา (ordinary least squares) มีค่าต่ำกว่าความเป็นจริงได้ Pindyck และ Rubinfeld (1998, P 302) กล่าวถึงกรณีนี้ว่าค่าประมาณของสโลปจากวิธีกำลังสองน้อยที่สุดแบบธรรมดา (ordinary least squares slope estimate) ที่ได้รับในกรณีนี้จะมีลักษณะ "เอนเอียง (biased)" เนื่องจากเป็นการประมาณค่าสโลปของการถดถอยที่แท้จริง (true regression slope) ต่ำกว่าความเป็นจริง และในทางตรงกันข้ามกันถ้าเรามีค่าสังเกต (observations) ซึ่งมีค่า x ที่มีลักษณะเกาะกลุ่มกันตรงกลางมากเกินไป(ซึ่งตรงกันข้ามกับกรณีแรกซึ่งเป็นกรณีปลายสุดหรือสุดโต่งเป็นจำนวนมากเกินไป) ค่าของ

สโลป (slope) ที่ประมาณค่าได้ก็จะมีลักษณะสูงเกินกว่าความเป็นจริง (overestimated) (ผู้อ่านที่สนใจการอธิบายโดยกราฟโปรดดู Pindyck และ Rubinfeld (1998, P 302)

จะเห็นได้ว่าแบบจำลองเชิงเส้นมีจุดอ่อนหลายประการด้วยกันดังได้กล่าวมาแล้วข้างต้น เพราะฉะนั้นต่อไปนี้เราจะมาพิจารณาทางเลือกอื่น เช่น แบบจำลองโพรบิท (probit model) ซึ่ง Glodberger (1964) เรียกว่าแบบจำลองวิเคราะห์แบบโพรบิท (probit analysis model) และแบบจำลองโลจิท (logit model) เป็นต้น

แบบจำลองโพรบิท (probit model)

จากแบบจำลองอย่างง่าย (1) เราสามารถเขียนให้อยู่ในรูปทั่วไปได้ดังนี้

$$y_i = x'_i\beta + u_i \qquad (17)$$

โดยที่ y_i = ตัวแปรตามแบบหุ่น (dummy dependent variable) ของค่าสังเกต i

x_i = k x 1 เวกเตอร์ของคุณลักษณะของค่าสังเกต i

β = k x 1 เวกเตอร์ของพารามิเตอร์

u_i = ค่าความคลาดเคลื่อนของค่าสังเกต i

แบบจำลอง (17) นี้เป็นแบบจำลองที่เราสังเกตค่า y_i ได้ ซึ่งแบบจำลอง (17) นี้ได้พัฒนามาจากการที่เราสมมุติว่า y^* มีความสัมพันธ์แบบถดถอย (regression relationship) ดังนี้

$$y^* = x'_i\beta + u_i \qquad (18)$$

ซึ่งในทางปฏิบัติแล้วค่า y^* จะเป็นตัวแปรที่เราไม่สามารถที่จะสังเกตได้ (unobservable) (maddala, 1983, P 22; johnston และ Dinardo, 1997, P 419) ซึ่ง Johnston และ Dinardo (1997, P 419) เรียก y^* ว่า "ตัวแปรที่ไม่ปรากฏให้เห็น (latent variable)" สิ่งที่เราสังเกตเห็นก็คือค่า y ซึ่งจะมีค่า 0 หรือ 1 ตามคำนิยาม (Maddala, 1983, P 22) หรือกฎ (rule) (Johnston และ Dinardo, 1997, P 419) ดังต่อไปนี้

$$\begin{aligned} y_i &= 1 \text{ ถ้า } y^* > 0 \\ &= 0 \text{ ในกรณีอื่น ๆ ที่ไม่ใช่ } y^* > 0 \end{aligned} \qquad (19)$$

โดยที่ $u_i \sim N(O, \sigma^2)$

และเนื่องจากแบบจำลองที่เรากำลังพิจารณาในบทนี้เป็นแบบจำลองความน่าจะเป็น (probability model) เพราะฉะนั้นแนวคิดของเราก็คือการแปลง (transform) $x'_i\beta$ ไปสู่ความน่าจะเป็น (probability) เพราะฉะนั้นสิ่งที่เราต้องการก็คือฟังก์ชั่น F ที่จะทำให้

$$prob(y_i = 1) = F(x'_i\beta)$$

ฟังก์ชั่น F ที่จะแปลง $x'_i\beta$ ให้อยู่ในระหว่าง 0 และ 1 ได้อย่างดีก็คือฟังก์ชั่นการแจกแจง (distribution function) หรือ ความหนาแน่นสะสม (cumulative density) (Johnston และ Dinardo, 1997, P 418) ซึ่งฟังก์ชั่นการแจกแจง (distribution function) นี้บางทีก็เรียกว่าฟังก์ชั่นการแจกแจงสะสม (cumulative distribution function) (Mendenhall และ Scheaffer, 1973, P 115) ตามสมการ (18) แบะ (19) $x'_i\beta$ จะไม่ใช้ $E(y_i/x_i)$ เหมือนอย่างที่เป็นในแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) แต่ $x'_i\beta$ ในกรณีนี้จะเท่ากับ $E(y_i^*/x_i)$ (Maddala, 1983, P 22)

จากสมการ (18) y_i^* (ภายใต้เงื่อนไขของ x) จะมีการแจกแจงแบบปกติ (normal distribution) แม้ว่า y_i^* (ซึ่งคือค่าที่ปรากฏของ y_i^* ตามคำนิยามหรือกฎ (19) จะไม่เป็นการแจกแจงแบบปกติก็ตามและจากคำนิยามหรือกฎ (19) เราสามารถที่เขียนได้ว่า

$$\begin{aligned} prob(y_i = 1) &= prob(y_i^* > 0) \\ &= prob(x'_iP = u_i > 0) \\ &= prob(u_i > -x'_i\beta) \\ &= prob(\frac{u_i}{\sigma} > -\frac{x_i\beta}{\sigma}) \end{aligned} \tag{20}$$

โดยที่ σ^2 คือความแปรปรวนของ u_i ดังได้กล่าวไว้แล้วข้างต้น การหารที่เกิดขึ้นในสมการ (20) จะทำให้เทอม u_i กลายเป็น u_i/σ ซึ่ง u_i/σ นี้มีการแจกแจง (distribution) เป็นแบบปกติ (standard normal distribution) (Johnston และ Dinardo, 1997, P 419) และจากสมการ (20) เราจะได้ว่า

$$\text{prob}(y_i = 1) = \text{prob}(\frac{u_i}{\sigma} > -\frac{x_i'\beta}{\sigma})$$
$$= \text{prob}(\frac{u_i}{\sigma} < \frac{x_i'\beta}{\sigma})$$
$$= \Phi\frac{x_i'\beta}{\sigma}$$

(Johnston และ Dinardo, 1997, P 874) (21)

โดยที่ $\Phi(.)$ คือการแจกแจงปกติมาตรฐาน (standard normal distribution) (Greene, 1997, P 874) ซึ่งสามารถเขียนสมการ (21) โดยเต็มรูปแบบได้ดังนี้

$$\text{prob}(y_i = 1) = \Phi\frac{x_i'\beta}{\sigma} = \int_{-\alpha}^{\frac{x_i\beta}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})\, dz \quad (22)$$

ซึ่งคือแบบจำลองโพรบิท (probit) การแปลงแบบการแจกแจงปกติมาตรฐาน (standard narmal distribution), $\Phi(.)$, เป็นการบังคับให้ความน่าจะเป็น (probability) อยู่ในช่วง 0 และ 1 นั่นคือ

$$\lim_{z \to +\alpha} \Phi(z) = 1$$

และ

$$\lim_{z \to -\alpha} \Phi(z) = 0$$

(Johnston และ Dinardo, 1997, P 418) (23)

จากสมการ (21)

$$\text{prob}(y_i = 1) = \Phi\frac{x_i'\beta}{\sigma}$$

สิ่งที่ตามมาก็คือ

$$\text{prob}(y_i = 1) = 1 - \text{prob}(y_i = 1)$$
$$= 1 - \Phi(\frac{x_i'\beta}{\sigma})$$

(Johnston และ Dinardo, 1997, P 419;Maddala, 1983, P 22 (24)

และถ้าตัวอย่างที่เราเลือกมีการแจกแจงที่เหมือนกันและเป็นอิสระต่อกัน (independently identical distribution, lid) และในกรณีนี้ค่า y ที่ได้มาหรือสังเกตมา (observed values ของ y) ก็คือค่าที่เกิดขึ้นจริงของกรรมวิธี binomial (binomial process) ด้วยความน่าจะเป็นตามสมการ (21) เราจะได้ความน่าจะเป็นร่วม (join probability) หรือ ฟังก์ชั่น ....(likelihood function) ดังนี้

$$L = \text{prob}(y_1 = 0).\text{prob}(y_2 = 0) \ldots \text{prob}(y_m = 0) .\text{prob}(y_{m+1}=1)\ldots\text{prob}(y_n = 1) \tag{25}$$

$$= \prod_{i=1}^{m} [1 - \Phi(\frac{X_i'\beta}{\sigma})]\Phi \prod_{i=m+1}^{n} (\frac{X_i'\beta}{\sigma}) \tag{26}$$

$$= \prod_{i=1}^{n} \Phi(\frac{X_i'\beta}{\sigma})^{y_i}\left[1 - \Phi(\frac{X_i'\beta}{\sigma})\right]^{1-y_i} \tag{27}$$

สมการ (27) ในรูปของลอการิซึม (logarithm) หรือ log – likelihood สามารถเขียนได้ดังนี้

$$\ln L = \sum_{i=1}^{n}\{y_i \ln[\Phi(\frac{X_i'\beta}{\sigma})] + (1-y_i).\ln[1-\Phi(\frac{X_i'\beta}{\sigma})]\} \tag{28}$$

$$= \sum_{y_i=0}\ln[1-\Phi(\frac{X_i'\beta}{\sigma})] + \sum_{y_i=1}\ln\Phi(\frac{X_i'\beta}{\sigma}) \tag{29}$$

(Johnston และ Dinardo, 1997, P 420; Greene,1997,p 882; Maddala, 1983, P 22) โปรดสังเกตุว่าค่า log – likelihood จะมีค่าสูงสุดไม่เกิน 0 เพราะว่า $0 \leq \Phi(.) \leq 1$ มีนัยว่า $\ln[1-\Phi(.)] \leq 0$ และ $\ln[-1\Phi(.)] \leq 0$ (johnston และ Dinardo, 1997, P 420) ลักษณะที่สำคัญอีกประการหนึ่งของ (likelihood function) ก็คือพารามิเตอร์ β และ σ จะปรากฏด้วยกันเสมอ เพราะฉะนั้นจะไม่สามารถหาค่าแยกออกมาต่างหากจากกันได้ สิ่งที่ได้ก็คือเรโซ β/σ เท่านั้น เพราะฉะนั้นจะเป็นการสะดวกที่จะ normalize σ ให้มีค่าเท่ากัน 1 เพื่อที่ว่าเราจะสามารถพูดถึง β อย่างเดียวได้

เงื่อนไขอันดับแรก (first – order) สำหรับการให้สมการ (28) มีค่าสูงสุด (maximization) ก็คือ

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left\{\frac{y_i\phi(.)}{\Phi(.)} + (1-y_i)\left[\frac{-\phi(.)}{1-\phi(.)}\right]\right\}x_i = 0 \tag{30}$$

$$= \sum_{y_i=0}\left[\frac{-\phi(\frac{x_i'\beta}{\sigma})}{1-\Phi(\frac{x_i'\beta}{\sigma})}\right]x_i + \sum_{y_i=1}\left[\frac{\phi(\frac{x_i'\beta}{\sigma})}{\Phi(\frac{x_i'\beta}{\sigma})}\right]x_i$$

$$= \sum_{i=1}^{n}\left[\frac{q_i\phi(\frac{q_i x_i'\beta}{\sigma})}{\Phi(\frac{q_i x_i'\beta}{\sigma})}\right]x_i$$

$$= \sum_{i=1}^{n}\lambda_i x_i = 0$$

(Greene,1997,P 882)

โดยที่ $q_i = 2y_{i-1}$

Φ_i = ฟังก์ชั่นความหนาแน่นของการแจกแจงปกติมาตรฐาน (standard normal density function)

สมการ (30) เป็นสมการที่ไม่ใช่เชิงเส้น (nonlinear) เพราะฉะนั้นการหาคำตอบก็จะต้องใช้วิธีการทำซ้ำ ๆ กัน (iterative method) สำหรับอนุพันธ์ครั้งที่ 2 (second derivatives) นั้นหามาได้โดยการใช้

$$\frac{d\phi(z)}{dz} = -z\phi(z)$$

ซึ่งจะได้

$$H = \frac{\partial^2 \ln L}{\partial \beta\, \partial \beta'} = \sum_{i=1}^{n} -\lambda_i(\lambda_i + x_i'\beta)\, x_i x_i' \tag{31}$$

ซึ่งมีค่าเป็น negative definite สำหรับทุกค่าของ β

สำหรับ asymtotic covariance matrix สำหรับตัวประมาณค่า (estimator) แบบ maximum likelihood นั้นหาได้จากการใช้ inverse ของ Hessian ที่คำนวณ ณ ค่าประมาณแบบ maximum likelihood นอกจากนี้ยังมีตัวประมาณค่า (estimators) อื่น ๆ อีก 2 ตัว ตัวประมาณค่า Berndt, Hall, Hall และ Hausman สามารถเขียนได้ดังนี้

$$B = \sum_i \lambda_i \mathrm{x}_i \mathrm{x}'_i$$

สำหรับตัวประมาณค่า (estimator) อีกตัวหนึ่งซึ่งอาศัยค่าคาดหมายของ Hessian ซึ่ง Greene (1997, P 884) กล่าวว่าจาก Amemiya (1981) จะได้

$$E\left[\frac{\partial^2 \ln L}{\partial \beta \, \partial \beta'}\right]_{probit} = \sum_{i=1}^{n} -\lambda_{0i} \lambda_{1i} \mathrm{x}_i \mathrm{x}'_i \qquad (32)\text{"}$$

สำหรับ logit model นั้นเราให้

$$\Phi(X_i'\beta) = \frac{1}{1 + e^{-x_i'\beta}}$$

3.3 แบบจำลองที่ใช้ในการศึกษา

เนื่องจากวัตถุประสงค์ที่สำคัญในการศึกษาครั้งนี้ คือการประมาณค่าพารามิเตอร์ของตัวแปรอิสระหรือปัจจัยที่คาดว่าจะมีผลต่อการตัดสินใจเลือกใช้บริการที่ปรึกษาทางด้านบัญชีจากบริษัทที่ปรึกษาที่ได้มาตรฐา ซึ่งแบบจำลองที่สามารถใช้ได้กับการศึกษาครั้งนี้ก็มีหลายแบบด้วยกัน เช่น แบบจำลองโทบิท (Tobit Model) แบบจำลองเส้นตรง (Linear Model) แบบจำลองโพรบิท (Probit Model) และแบบจำลองโลจิท (Logit Model) เป็นต้น สำหรับแบบจำลองโทบิทนั้น ไม่สามารถนำมาใช้ในการศึกษานี้ได้ เนื่องจากตัวแปรตามในแบบจำลองจะมีค่าแบบไม่ต่อเนื่อง ดังนั้น แบบจำลองที่สามารถใช้ในการศึกษานี้จึงมีเพียงแบบจำลองเส้นตรง แบบจำลองโพรบิท และ

แบบจำลองโลจิทเท่านั้น แต่อย่างไรก็ตาม เนื่องจากแบบจำลองแบบเส้นตรงมีข้อบกพร่อง คือ ค่ากะประมาณหรือค่าพยากรณ์ของตัวแปรตามที่ได้จากแบบจำลองอาจมีค่าเกินกว่า 1 หรือ น้อยกว่า 0 และแบบจำลองแบบโพรบิทนั้นมีข้อโต้แย้งว่าแนวโน้มทฤษฎีสนับสนุนค่อนข้างจำกัดและการคำนวณค่อนข้างยุ่งยาก ดังนั้น การศึกษานี้จึงเลือกที่จะใช้แบบจำลองโลจิท ซึ่งกำหนดแบบจำลองไว้ดังนี้

ให้แบบจำลองเป็น

$$y_i^* = X_i'\beta + u_i$$

ในทางปฏิบัติ y_I^* ไม่สามารถสังเกตได้ (unobservable) สิ่งที่เราสังเกตมาได้ก็คือ ตัวแปรหุ่น (dummy variable) y ซึ่งนิยามได้ดังนี้

$$y = 1 \quad \text{ถ้า } y_I^* > 0$$

$$y = 0 \quad \text{อื่น ๆ}$$

$$= \prod_{y_i=0}[1-\Phi(X_i'\beta)]\prod_{y_i=1}\Phi(X_i'\beta) \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

และจาก Likelihood function

$$\text{Prob}\,(Y_1 = y_1, Y_2 = y_2, \ldots Y_n = y_n)$$

ซึ่งสามารถเขียนในรูปที่กระทัดรัดได้ ดังนี้

$$L = \prod_{i=1}^{n}[\Phi(X_i'\beta)]^{y_i}[1-\Phi(X_i'\beta)]^{1-y_i} \ldots\ldots\ldots\ldots(2)$$

โดยที่

$$\Phi(X_i'\beta) = \frac{1}{1+e^{-x_i'\beta}}$$

(Johnston,1984)

$L =$ ความน่าจะเป็นที่ผู้ประกอบการในจังหวัดเชียงใหม่และในนิคมอุตสาหกรรมภาคเหนือจะเลือกใช้บริการจากกลุ่มบริษัทที่ปรึกษาทางด้านบัญชีที่ได้มาตรฐาน

X_1 = 1 ถ้าผู้ประกอบการเป็นบริษัทจำกัดที่ตั้งอยู่ในนิคมอุตสาหกรรมภาคเหนือ
0 ถ้าผู้ประกอบการไม่ใช่บริษัทจำกัดที่ตั้งอยู่ในนิคมอุตสาหกรรมภาคเหนือ

X_2 = 1 ถ้าผู้ถือหุ้นของผู้ประกอบการนั้นเป็นคนไทยล้วน
0 ถ้าผู้ถือหุ้นของผู้ประกอบการนั้นไม่ใช่คนไทยล้วน

X_3 = 1 ถ้าผู้ถือหุ้นของผู้ประกอบการนั้นเป็นคนญี่ปุ่นล้วน เกาหลีล้วน คนไทยและ/หรือคนญี่ปุ่นหรือเกาหลี
0 ถ้าผู้ถือหุ้นของผู้ประกอบการนั้นไม่ใช่คนญี่ปุ่นล้วน เกาหลีล้วน คนไทยและ/หรือคนญี่ปุ่นหรือเกาหลี

X_4 = ระดับการศึกษาของผู้มีอำนาจในการตัดสินใจของผู้ประกอบการรายนั้นๆ (จำนวนปีที่ศึกษา)

X_5 = ประสบการณ์ในการทำงานของผู้มีอำนาจในการตัดสินใจ (จำนวนปีที่ทำงาน)

X_6 = 1 ผลประกอบการของกิจการถ้ามีกำไร
0 ผลประกอบการของกิจการถ้าประสบผลขาดทุน

และโดยที่แบบจำลองนี้เป็นแบบจำลองโลจิท ดังนั้น จึงจะใช้วิธีการประมาณพารามิเตอร์โดยวิธีภาวะความน่าจะเป็นสูงสุด (Maximum-Likelihood Estimation) แบบจำลองดังกล่าวข้างต้น จะนำไปใช้ประมาณค่าพารามิเตอร์ โดยแสดงถึงค่าพารามิเตอร์ที่จะมีผลกระทบต่อความน่าจะเป็นที่ผู้ประกอบการจะตัดสินใจเลือกใช้บริการที่ปรึกษาทางด้านบัญชีโดยแบ่งบริษัทที่ปรึกษาทางด้านบัญชีออกเป็น 2 กลุ่ม คือ กลุ่มที่ได้รับการยอมรับว่าได้มาตรฐาน และกลุ่มที่ไม่ได้รับการยอมรับว่าได้มาตรฐาน ซึ่งในแบบจำลองนี้จะกำหนดให้

L = ความน่าจะเป็นที่ผู้ประกอบการในจังหวัดเชียงใหม่และ ในนิคมอุตสาหกรรมภาคเหนือจะตัดสินใจเลือกใช้บริการจากกลุ่มบริษัทที่ปรึกษาทางด้านบัญชีที่ได้มาตรฐานโดยจะมีค่าเท่ากับ 1 ถ้าผู้ประกอบการตัวอย่างเลือกใช้บริการจากกลุ่มบริษัทที่ปรึกษาทางด้านบัญชีที่ได้มาตรฐานและจะมีค่าเท่ากับ 0 ถ้าผู้ประกอบการตัวอย่างไม่ใช่บริการจากกลุ่มบริษัทที่ปรึกษาที่ได้มาตรฐาน