ภาคผนวก ก

วิธีการทางเศรษฐมิติเกี่ยวกับแบบจำลองสองทางเลือก (Binary choice model)

แบบจำลองสมการถดถอยโดยทั่วไปจะสมมติว่าตัวแปรตามเป็นข้อมูลเชิงปริมาณ (quantitative data) ส่วนตัวแปรอธิบายสามารถเป็นได้ทั้งข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ (qualitative data) แต่ในความเป็นจริงตัวแปรตามสามารถมีลักษณะข้อมูลเชิงคุณภาพได้ด้วย ซึ่งในแบบจำลองที่ตัวแปรตาม , y , มีลักษณะข้อมูลเชิงปริมาณ สิ่งที่ต้องการคือ การประมาณค่าค่าคาดหวัง (expected value) หรือค่าเฉลี่ย เมื่อกำหนดค่าตัวแปรอธิบาย , x , มาให้ ในขณะที่แบบจำลองที่ y มีลักษณะข้อมูลเชิงคุณภาพ สิ่งที่ต้องการคือ การหาค่าความน่าจะเป็นที่สิ่งใดสิ่งหนึ่งจะเกิดขึ้น (probability of something happening)

พิจารณาสมการถดถอย แสดงโดย

$$y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{(ก.1)}$$

โดยที่ X_i คือ กลุ่มของตัวแปรอธิบาย y คือ ตัวแปรข้อมูลแบบสองทางเลือก (binary choice variable) โดย y=1 เมื่อหน่วยผลิตตัดสินใจทำกิจกรรมการวิจัยและพัฒนา และเท่ากับ 0 เมื่อตัดสินใจไม่ทำการวิจัยและพัฒนา

สมการที่ ก.1 ดูเหมือนว่าเป็นแบบจำลองสมการถดถอยเชิงเส้นทั่วไป (linear regression model) แต่เนื่องจากตัวแปรตามมีลักษณะข้อมูลแบบสองทางเลือก จึงเรียกแบบจำลองลักษณะนี้ว่า “แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model)” ทั้งนี้ ค่าคาดหวังแบบมีเงื่อนไข (conditional expectation) ของ y_i เมื่อกำหนดค่า X_i , $E\left(y_i \middle| X_i\right)$, สามารถอธิบายได้ว่าคือ ความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ที่เหตุการณ์อาจจะเกิดขึ้นเมื่อกำหนดค่า X_i มาให้ , $\Pr\left(y_i = 1 \middle| X_i\right)$, ซึ่งถ้า P_i คือ ความน่าจะเป็นที่ y_i=1 (เหตุการณ์เกิดขึ้น) และ $(1- P_i)$ คือ ความน่าจะเป็นที่ y_i=0 (เหตุการณ์ไม่เกิดขึ้น) ตัวแปร y_i จะมีการแจกแจงดังนี้

y_i	probability
0	$1-P_i$
1	P_i
Total	1

จะเห็นได้ว่า y_i มีการแจกแจงแบบ Bernoulli probability distribution ซึ่งจากนิยามของความคาดหวังทางคณิตศาสตร์ (mathematical expectation) จะได้ว่า

$$E(y_i) = 0(1-P_i) + 1(P_i) = P_i$$

เนื่องจาก $E\left(y_i \middle| X_i\right) = \beta_1 + \beta_2 X_i$ เมื่อทำให้เท่ากันจะได้ว่า

$$E\left(y_i \middle| X_i\right) = \beta_1 + \beta_2 X_i = P_i \qquad \text{(ก.2)}$$

และเนื่องจากค่าความน่าจะเป็น P_i มีค่าระหว่าง 0 ถึง 1 จึงทำให้ค่าคาดหวังแบบมีเงื่อนไข (conditional expectation) หรือความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) มีค่าอยู่ระหว่าง 0 ถึง 1 ด้วย , $0 \le E\left(y_i \middle| X_i\right) \le 1$

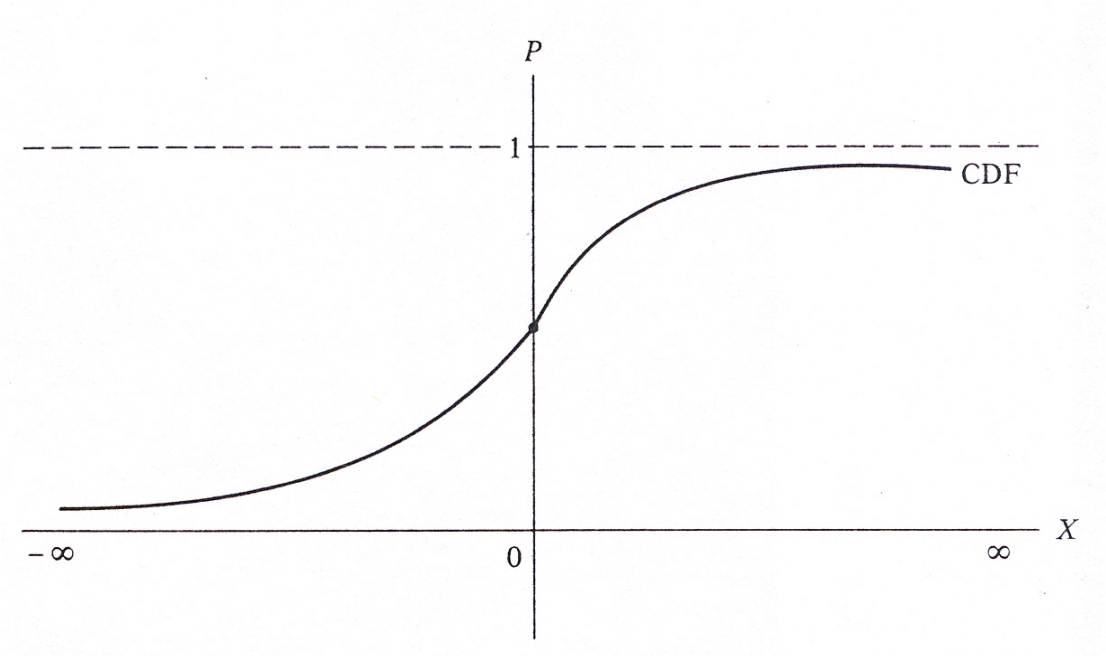
สำหรับการแปลความหมายของสัมประสิทธิ์ที่ได้จากแบบจำลองสามารถแปลความหมายได้โดยตรง เมื่อกำหนดให้ค่าของตัวแปรอื่นๆคงที่ ทุกๆการเปลี่ยนแปลงในค่าของ X_i จะส่งผลให้ค่าความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นเปลี่ยนแปลงท่ากับ β_i อย่างไรก็ตาม ถึงแม้ว่าแบบจำลองความน่าจะเป็นเชิงเส้นจะมีความง่ายต่อการนำไปใช้ แต่สมมติฐานของแบบจำลองหลายๆอย่างไม่ได้เป็นไปตามที่คาดไว้ เช่น การที่ค่าความคลาดเคลื่อนมีการแจกแจงแบบไม่ปกติ (non-normal error) ค่าความแปรปรวนไม่คงที่ (non-constant error variance) และรูปแบบสมการ (functional form) ที่มีลักษณะเชิงเส้นกับตัวแปร X เป็นต้น ซึ่งปัญหาที่ส่งผลให้แบบจำลองนี้ไม่เป็นที่น่าสนใจเท่าใดนัก คือ ปัญหาในเรื่องของการเพิ่มขึ้นในลักษณะเชิงเส้นกับตัวแปร X (increase linearly with X) หรือกล่าวอีกนัยหนึ่งคือ การที่ค่าผลกระทบส่วนเพิ่ม (marginal effect) ของ X มีค่าคงที่ตลอด ซึ่งเป็นสิ่งที่ไม่สมเหตุสมผลเท่าใดนัก เช่น สมมติว่า เมื่อรายได้เพิ่มขึ้น 1 หน่วย ($1,000) จะส่งผลให้ความน่าจะเป็นในการมีบ้านเป็นของตนเองเพิ่มขึ้นในสัดส่วนคงที่เท่ากับ 0.1 นั่นหมายถึง หากระดับของรายได้เพิ่มขึ้นเป็น $8,000, $10,000 หรือ $20,000 ก็จะมีความน่าจะเป็นเพิ่มขึ้นในสัดส่วนเดิมตลอด ซึ่งดูไม่ค่อยสมเหตุสมผลเท่าใดนัก โดยในความเป็นจริงค่าความน่าจะเป็น P_i มีความสัมพันธ์ลักษณะไม่เป็นเชิงเส้นกับตัวแปร X_i คือ เมื่อครอบครัวมีรายได้น้อยมากมักจะไม่มีบ้านเป็นของตนเอง แต่เมื่อมีรายได้เพิ่มขึ้นในระดับสูงก็จะมีแนวโน้มที่จะมีบ้านเป็นของตนเองสูงขึ้น เป็นต้น

ดังนั้น จากข้อจำกัดที่กล่าวมาจึงเกิดแนวคิดแบบจำลองขึ้นมาใหม่ โดยแบบจำลองดังกล่าวมีความน่าจะเป็นอยู่ระหว่าง 0 ถึง 1 และมีความสัมพันธ์ในลักษณะไม่เป็นเชิงเส้นกับตัว

แปร X ทั้งนี้ ฟังก์ชันการแจกแจงความน่าจะเป็นสะสม (cumulative distribution function : CDF) ของตัวแปรสุ่มดูเหมือนจะสามารถตอบสนองต่อเงื่อนไขดังกล่าวได้ เนื่องจากฟังก์ชันดังกล่าวมีลักษณะการกระจายตัวของข้อมูลแบบ S-shape ดังแสดงในภาพที่ ก.1 ซึ่งฟังก์ชันการแจกแจงความน่าจะเป็นสะสมของตัวแปรสุ่ม X คือ ฟังก์ชันที่แสดงถึงค่าความน่าจะเป็นสะสมที่ X จะมีค่าน้อยกว่าค่าใดค่าหนึ่ง (X_0) เขียนเป็นสมการได้ว่า $F\left(x = x_0\right) = P(x \le x_0)$ ดังนั้น จึงเป็นการง่ายที่จะใช้ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมในการสร้างแบบจำลองสมการถดถอยเมื่อตัวแปรตามมีลักษณะข้อมูลแบบสองทางเลือก โดยฟังก์ชันการแจกแจงความน่าจะเป็นสะสมที่นิยมใช้โดยทั่วไปคือ ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมแบบโลจิสติก (logistic cumulative distribution function) และฟังก์ชันการแจกแจงความน่าจะเป็นสะสมแบบปกติ (normal cumulative distribution function) โดยเรียกแบบจำลองที่ใช้ฟังก์ชันการแจงแจงแบบนี้ว่าแบบจำลองโลจิต (logit model) และแบบจำลองโพรบิท (probit หรือ normit model) ตามลำดับ

ภาพที่ ก.1

การแจกแจงความน่าจะเป็นสะสม (cumulative distribution function : CDF)



ที่มา : Gujarati (2003) หน้า 594

1. <u>แบบจำลองโลจิต (Logit model)</u>

จากแบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model) ซึ่งมีความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) คือ

$$P_i = E\left(y = 1 \middle| X_i\right) = \beta_1 + \beta_2 X_i$$

โดยที่ X_i คือ กลุ่มของตัวแปรอธิบาย และ y คือ ตัวแปรตามซึ่งมีลักษณะข้อมูลแบบสองทางเลือก ทั้งนี้ จากที่ได้กล่าวมาแล้วว่า ในการอธิบายถึงพฤติกรรมของตัวแปรตามที่มีลักษณะข้อมูลแบบสองทางเลือกจำเป็นต้องเลือกใช้ฟังก์ชันการแจกแจงความน่าจะเป็นสะสม (CDF) ที่มีความเหมาะสม ซึ่งในแบบจำลองโลจิตเลือกใช้ฟังก์ชันการแจกแจงความน่าจะเป็นสะสมแบบโลจิสติก (logistic cumulative distribution function) ดังนี้

$$P_i = E\left(y = 1 \middle| X_i\right) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$$

หรือเขียนใหม่ให้อยู่ในรูปของ

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}} \qquad \text{(ก.3)}$$

โดยที่ $Z_i = \beta_1 + \beta_2 X_i$ และเรียกสมการที่ ก.3 ว่า ฟังก์ชันการแจกแจงแบบโลจิสติก (logistic distribution function) ทั้งนี้ ค่า Z_i มีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ ทำให้ค่า P_i มีค่าอยู่ระหว่าง 0 ถึง 1[1] นอกจากนี้ค่า P_i ยังมีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X_i ซึ่งสอดคล้องกับเงื่อนไขที่ได้กล่าวมาแล้ว อย่างไรก็ตาม จะเห็นได้ว่านอกจาก P_i จะมีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X_i แล้ว ยังมีลักษณะไม่เป็นเชิงเส้นกับพารามิเตอร์ β อีกด้วย ทำให้ไม่สามารถใช้ขั้นตอนของวิธีกำลังสองน้อยที่สุด (ordinary least square : OLS) ในการประมาณค่าพารามิเตอร์ได้ แต่ปัญหานี้สามารถแก้ไขได้ หากทำให้สมการที่ ก.3 มีลักษณะแบบเชิงเส้น แสดงโดย

---

[1] เมื่อ $Z_i \to +\infty$ แล้ว e^{-Z_i} จะมีค่าเข้าใกล้ 0 และเมื่อ $Z_i \to -\infty$, e^{-Z_i} จะมีค่าเข้าสู่อนันต์ โดยที่ $e \approx 2.71828$

จาก
$$P_i = \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{1+e^{Z_i}}$$
$$1-P_i = 1-\frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{Z_i}} \qquad \text{(ก.4)}$$

(3)/(4)
$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} \qquad \text{(ก.5)}$$
$$= e^{(\beta_1+\beta_2 X_i)} = e^{\beta_1} \cdot e^{\beta_2 X_i}$$

ทั้งนี้ เราเรียก $\frac{P_i}{1-P_i}$ ว่า อัตราส่วนความเป็นไปได้ (odds ratio) แสดงถึง อัตราส่วนของความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นต่อความน่าจะเป็นที่เหตุการณ์จะไม่เกิดขึ้น เช่น หาก P_i=0.8 หมายความว่า มีโอกาสที่เหตุการณ์จะเกิดขึ้นเท่ากับ 4 ต่อ 1 ทั้งนี้ ในการแปลความหมายของการเปลี่ยนแปลงค่าของ X ที่มีต่ออัตราส่วนความเป็นไปได้ สามารถอธิบายได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงของค่า X 1 หน่วย จะทำให้อัตราส่วนความเป็นไปได้เปลี่ยนแปลง e^{β_2} เท่า

หากเราใส่ natural logarithm ในสมการ ก.5 จะได้ว่า

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i \qquad \text{(ก.6)}$$

โดยที่ L_i = log of odds ratio ซึ่งจะเห็นได้ว่าสมการดังกล่าวมีลักษณะเชิงเส้นทั้งกับตัวแปร X_i และพารามิเตอร์ และเรียกสมการ ก.6 ว่าคือ แบบจำลองโลจิต (logit model) โดยมีสมการที่ใช้ในการประมาณค่าแบบจำลองโลจิต คือ

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i + u_i \qquad \text{(ก.7)}$$

ทั้งนี้ กำหนดให้ P_i=1 ถ้าบริษัทตัดสินใจทำการวิจัยและพัฒนา และเท่ากับ 0 เมื่อบริษัทตัดสินใจไม่ทำการวิจัยและพัฒนา ซึ่งหากเราแทนค่า P_i โดยตรงในสมการโลจิต จะได้ว่า

$$L_i = \ln\left(\frac{1}{0}\right) \qquad \text{หากบริษัทตัดสินใจทำการวิจัยและพัฒนา}$$

$$L_i = \ln\left(\frac{0}{1}\right) \qquad \text{หากบริษัทตัดสินใจไม่ทำการวิจัยและพัฒนา}$$

จะเห็นได้ว่า การคำนวณดังกล่าวไม่สามารถทำได้ ดังนั้น เราจึงไม่สามารถใส่ค่าได้โดยตรงและใช้วิธีกำลังสองน้อยที่สุด (ordinary least square : OLS) ในการประมาณค่าได้ ซึ่งในกรณีเช่นนี้จำเป็นต้องอาศัยวิธีการประมาณค่าแบบ maximum likelihood ในการประมาณค่าพารามิเตอร์ โดยเริ่มจาก สมมติว่าสิ่งที่เราสนใจคือ การประมาณค่าความน่าจะเป็นที่บริษัทจะตัดสินใจทำการวิจัยและพัฒนา และกำหนดให้ความน่าจะเป็นนี้คำนวณได้จากฟังก์ชันการแจกแจงแบบโลจิสติก แสดงโดย

จาก $$P_i = \frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}}$$

เนื่องจากเราไม่สามารถสังเกตค่า P_i ที่แท้จริงได้ สังเกตได้เพียงแต่ค่า y=1 หากบริษัทตัดสินใจทำการวิจัยและพัฒนา และ y=0 หากบริษัทตัดสินใจไม่ทำการวิจัยและพัฒนา ซึ่งแต่ละค่า y_i ที่สังเกตได้ คือ ตัวแปรสุ่มแบบ Bernoulli เขียนได้ดังนี้

$$\Pr(y_i = 1) = P_i$$
$$\Pr(y_i = 0) = 1 - P_i$$

หากสมมติว่ามีตัวแปรสุ่ม n ตัวอย่าง และกำหนดให้ $f_i(y_i)$ คือ ความน่าจะเป็นที่ y_i=1 หรือ 0 ซึ่ง joint probability density function ของ $y_1, y_2, \ldots, y_n$ คือ $f(y_1, y_2, \ldots, y_n)$ แสดงโดย

$$f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} P_i^{y_i}\left(1-P_i\right)^{1-y_i} \qquad \text{(ก.8)}$$

ทั้งนี้เราสามารถเขียน joint probability density function ว่าคือ ผลคูณของฟังก์ชันความหนาแน่น (density function) ของแต่ละตัวอย่าง y_i ที่เป็นอิสระจากกัน และมีฟังก์ชันความหนาแน่นแบบโลจิสติกเหมือนกัน โดยเรียกสมการ ก.8 ว่า likelihood function

เนื่องจากสมการที่ ก.8 มีความไม่สะดวกต่อการจัดการ เราจึงแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) และเรียกฟังก์ชันแบบนี้ว่า log likelihood function แสดงโดย

$$\begin{aligned}\ln L = \ln f(y_1, y_2, \ldots, y_n) &= \sum_{i=1}^{n}\left[y_i \ln P_i + (1-y_i)\ln(1-P_i)\right] \\ &= \sum_{i=1}^{n}\left[y_i \ln P_i - y_i \ln(1-P_i) + \ln(1-P_i)\right] \\ &= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{P_i}{1-P_i}\right)\right] + \sum_{i=1}^{n}\ln(1-P_i)\end{aligned} \qquad \text{(ก.9)}$$

แทนค่า $\ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i$ และ $1-P_i = \frac{1}{1+e^{\beta_1+\beta_2 X_i}}$ ลงในสมการ ก.9 จะได้ว่า

$$\begin{aligned}\ln L = \ln f(y_1, y_2, \ldots, y_n) &= \sum_{i=1}^{n} y_i\left(\beta_1 + \beta_2 X_i\right) + \sum_{i=1}^{n}\ln\left(\frac{1}{1+e^{\beta_1+\beta_2 X_i}}\right) \\ &= \sum_{i=1}^{n} y_i\left(\beta_1 + \beta_2 X_i\right) - \sum_{i=1}^{n}\ln\left(1+e^{\beta_1+\beta_2 X_i}\right)\end{aligned} \qquad \text{(ก.10)}$$

จะเห็นได้ว่า log likelihood function เป็นฟังก์ชันของพารามิเตอร์ β_1 และ β_2 เนื่องจากค่า X_i เป็นค่าที่ทราบอยู่แล้ว ภายใต้เงื่อนไขลำดับที่หนึ่ง (first order condition) ของสมการที่ ก.10 เพื่อหาค่าพารามิเตอร์ที่ไม่ทราบค่า จะได้ว่า

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_1} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\left(\frac{e^{\beta_1}}{1+e^{Z_i}}\right) = 0 \\ \frac{\partial \ln L}{\partial \beta_2} &= \sum_{i=1}^{n} y_i X_i - \sum_{i=1}^{n}\left(\frac{e^{\beta_2 X_i}}{1+e^{Z_i}}\right) X_i = 0\end{aligned} \qquad \text{(ก.11)}$$

ทำการแก้สมการ ก.11 เพื่อหาค่าประมาณ $\hat{\beta}_1$ และ $\hat{\beta}_2$ ทั้งนี้ การประมาณแบบ maximum likelihood มีจุดประสงค์เพื่อทำให้ likelihood function มีค่าสูงสุด ซึ่งก็คือการประมาณค่าพารามิเตอร์ที่ไม่ทราบค่า เพื่อทำให้เซทของกลุ่มตัวอย่างที่เราสังเกตได้มีโอกาสเกิดขึ้นได้มากที่สุด อย่างไรก็ตาม ควรทำความเข้าใจว่า ในขั้นตอนการคำนวณจะมีลักษณะไม่เป็นเชิงเส้นกับพารามิเตอร์สูง (highly nonlinear in parameter) และไม่สามารถแก้ปัญหาได้อย่างชัดเจน

(no explicit solution) ทำให้จำเป็นต้องใช้วิธีการประมาณค่าสมการแบบไม่เป็นเชิงเส้นเพื่อให้ได้คำตอบที่เป็นตัวเลข

จากการประมาณค่าพารามิเตอร์ที่ไม่ทราบค่าจะได้สมการประมาณค่าโลจิต ดังนี้

$$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \text{(ก.12)}$$

สำหรับการแปลความหมาย อธิบายได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงเพิ่มขึ้นของค่า X_i 1 หน่วย จะส่งผลให้ log of odds เพิ่มขึ้น $\hat{\beta}_2$ หน่วย หากทำการ antilog สมการ ก.12 จะได้ว่า

$$\frac{\hat{P}_i}{1-\hat{P}_i} = e^{\hat{\beta}_1+\hat{\beta}_2 X_i} = e^{\hat{\beta}_1} \cdot e^{\hat{\beta}_2 X_i} \qquad \text{(ก.13)}$$

ในการแปลความหมายอธิบายได้ว่า อธิบายได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงเพิ่มขึ้นของค่า X_i 1 หน่วย จะส่งผลให้อัตราส่วนความเป็นไปได้เพิ่มขึ้น $e^{\hat{\beta}_2}$ เท่า หรือ $\left(e^{\hat{\beta}_2}-1\right)\times 100\%$ นอกจากนี้ ยังสามารถคำนวณหาอัตราการเปลี่ยนแปลงของความน่าจะเป็น (rate of change of probability) หรือผลกระทบส่วนเพิ่ม (marginal effect) โดยที่

$$\begin{aligned}
\frac{dE\left(y|X_i\right)}{dX_i} &= \frac{dP_i}{dX_i} \\
&= \frac{d\left(\frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}}\right)}{dX_i} \\
&= \frac{e^{-\beta_2 X_i}\cdot\beta_2}{\left(1+e^{-\beta_2 X_i}\right)^2} \\
&= \frac{1}{\left(1+e^{-\beta_2 X_i}\right)}\cdot\frac{e^{-\beta_2 X_i}}{\left(1+e^{-\beta_2 X_i}\right)}\cdot\beta_2 \\
&= \frac{1}{\left(1+e^{-\beta_2 X_i}\right)}\cdot\frac{1}{\left(1+e^{\beta_2 X_i}\right)}\cdot\beta_2
\end{aligned}$$

จะได้ว่า

$$\frac{dE\left(y|X_i\right)}{dX_i} = \frac{dP_i}{dX_i} = \beta_2 P_i\left(1-P_i\right) \qquad \text{(ก.14)}$$

สมการที่ ก.14 แสดงถึง อัตราการเปลี่ยนแปลงในความน่าจะเป็นเนื่องจากการเปลี่ยนแปลงใน X ซึ่งไม่เพียงแต่ประกอบด้วย β_2 เท่านั้น ยังประกอบด้วยค่าของระดับความน่าจะเป็นอีกด้วย ซึ่งระดับของความน่าจะเป็นจะเปลี่ยนแปลงตามค่า X_i โดยผลของการเปลี่ยนแปลงในค่า X_i ที่มีต่อ P_i จะมีค่าสูงสุดเมื่อ P_i=0.5 และจะมีค่าต่ำสุดเมื่อ P เข้าใกล้ค่า 0 หรือ 1 ซึ่งอัตราการเปลี่ยนแปลงของความน่าจะเป็นที่ได้จะมีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X ทั้งนี้ในการแปลความหมาย อธิบายได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงเพิ่มขึ้นของค่า X_i 1 หน่วย จะส่งผลให้ค่าความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นเปลี่ยนแปลงเท่ากับ $\hat{\beta}_2\hat{P}_i\left(1-\hat{P}_i\right)$

สำหรับการวัดค่าความน่าเชื่อถือของแบบจำลอง (goodness of fit) สำหรับแบบจำลองที่ตัวแปรตามมีลักษณะข้อมูลแบบสองทางเลือกมักจะนิยมวัดจากค่าที่มีความคล้ายคลึงกับ R^2 โดยเรียกว่า pseudo R^2 ทั้งนี้ เนื่องจากการวัดโดยใช้ R^2 อาจจะให้ค่าที่ต่ำกว่า 1 มาก ซึ่ง Aldrich and Nelson ยืนยันว่า ควรจะหลีกเลี่ยงการใช้ R^2 ในการสรุปผลความน่าเชื่อถือของแบบจำลอง (goodness of fit) สำหรับแบบจำลองที่ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ อนึ่ง การประเมินความน่าเชื่อถือของแบบจำลองโดยใช้ pseudo R^2 มีความหลากหลายมาก ทั้งนี้ จะยกตัวอย่างค่า pseudo R^2 ที่คำนวณได้จากโปรแกรม Eviews ซึ่งก็คือ McFadden R^2 โดยนิยามว่า

$$R^2_{McF} = 1 - \frac{LLF_{ur}}{LLF_r}$$

โดยที่ LLF_{ur} คือ unrestricted log likelihood function ซึ่งตัวแปรทุกตัวจะถูกรวมอยู่ในแบบจำลอง ส่วน LLF_r คือ restricted log likelihood function จะมีเฉพาะค่าจุดตัดแกน (intercept) เท่านั้นที่ถูกรวมอยู่ในแบบจำลอง โดยหลักการแล้ว LLF_{ur} จะเปรียบได้กับ RSS ขณะที่ LLF_r จะมีค่าเท่ากันกับ TSS ในแบบจำลองสมการถดถอย ทั้งนี้ R^2_{McF} มีค่าอยู่ระหว่าง 0 ถึง 1 อย่างไรก็ตาม ไม่ควรจะให้ความสำคัญมากเกินไปกับค่าความน่าเชื่อถือของแบบจำลองที่ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ

2. <u>แบบจำลองโพรบิท (Probit model)</u>

ในการอธิบายถึงพฤติกรรมของตัวแปรตามที่มีลักษณะข้อมูลแบบสองทางเลือก (binary choice dependent variable) จำเป็นต้องเลือกใช้ฟังก์ชันการแจกแจงความน่าจะเป็น

สะสม (CDF) ที่มีความเหมาะสม โดยที่ผ่านมาได้กล่าวถึงแบบจำลองโลจิตซึ่งใช้ฟังก์ชันการแจกแจงสะสมแบบโลจิสติก ทั้งนี้ ไม่ได้มีเพียงแต่ฟังก์ชันการแจกแจงสะสมแบบโลจิสติกเท่านั้นที่เป็นที่นิยม ยังมีฟังก์ชันการแจกแจงสะสมแบบปกติ (normal CDF) ที่สามารถนำมาใช้ได้อีกด้วย โดยเรียกแบบจำลองที่ใช้ฟังก์ชันการแจกแจงสะสมแบบปกติว่า แบบจำลองโพรบิท (probit model) และบางครั้งยังเรียกว่า แบบจำลอง normit model ได้อีกด้วย โดยการอธิบายแบบจำลองโพรบิทจะอาศัยทฤษฎีอรรถประโยชน์ (utility theory) หรือพฤติกรรมการเลือกอย่างมีเหตุผล ซึ่งพัฒนาโดย McFadden (1973) โดยอธิบายว่า หน่วยผลิตจะเปรียบเทียบระหว่างผลประโยชน์ส่วนเพิ่ม (marginal benefit) และต้นทุนส่วนเพิ่ม (marginal cost) ในการตัดสินใจว่าจะทำกิจกรรมวิจัยและพัฒนาหรือไม่ ซึ่งกำหนดให้ความแตกต่างระหว่างกำไรและต้นทุนเป็นค่าที่ไม่สามารถสังเกตได้ , y^* แสดงโดย

$$y^* = X'\beta + \varepsilon$$

โดยที่ X คือ เวกเตอร์ของตัวแปรอธิบาย, $X'\beta$ คือ index function และ ε มีการแจกแจงแบบปกติ , $\varepsilon \sim N(\mu, \sigma)$ ทั้งนี้ หน่วยผลิตจะตัดสินใจทำการวิจัยและพัฒนาหากกำไรสุทธิที่แท้จริงมีค่ามากกว่าค่าวิกฤติ (threshold) ค่าหนึ่ง (ในที่นี้สมมติให้เท่ากับ 0) ซึ่งหาก $y^* > 0$ หน่วยผลิตจะตัดสินใจทำการวิจัยและพัฒนา แต่เนื่องจากเราไม่สามารถสังเกตค่ากำไรสุทธิที่แท้จริงได้ ทำได้เพียงแต่สังเกตค่า y=1 หากบริษัทตัดสินใจทำการวิจัยและพัฒนา และ y=0 หากบริษัทตัดสินใจไม่ทำการวิจัยและพัฒนา ดังนั้นค่าที่สังเกตได้จึงมีลักษณะดังนี้

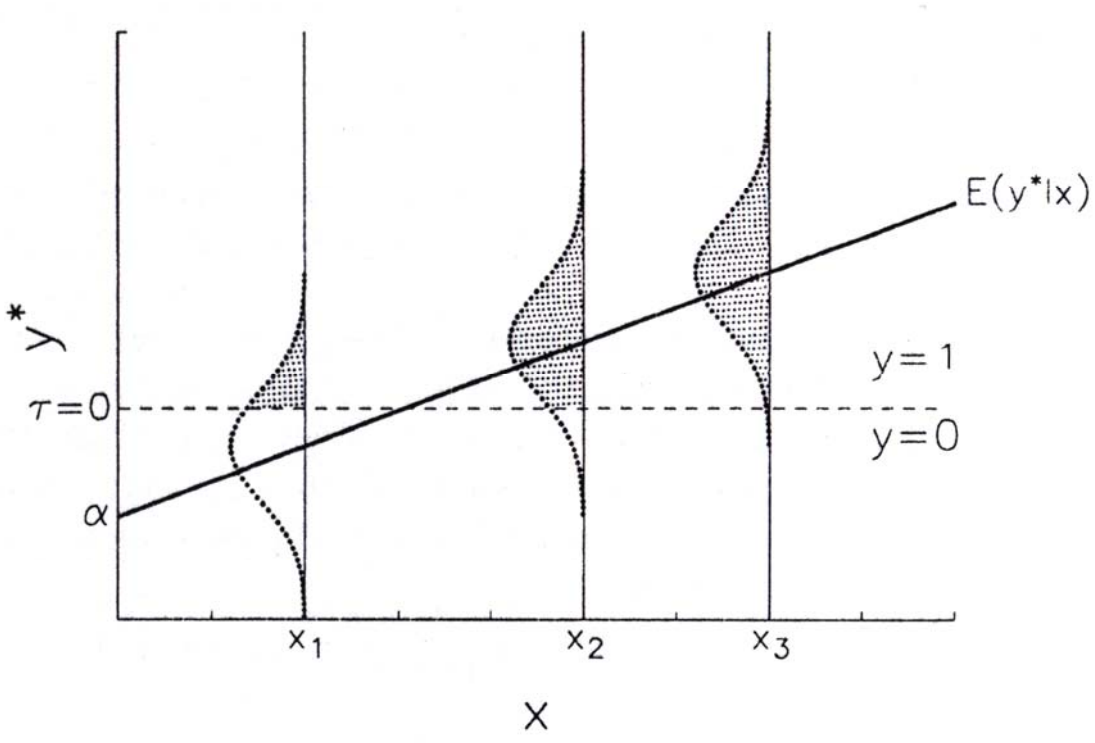
y=1 ถ้า $y^* > 0$

y=0 ถ้า $y^* \leq 0$

ในการอธิบายความเชื่อมโยงระหว่างตัวแปรแฝง y^* กับค่าที่สังเกตได้ ,y, สามารถอธิบายดังแสดงในภาพที่ ก.2 โดยที่ y^* อยู่ในแกนตั้งและค่าวิกฤติ (threshold), τ , ซึ่งแสดงโดยเส้นประ จะเห็นได้ว่าการกระจายตัวของค่า y^* มีลักษณะแบบระฆังคว่ำ (bell shape curve) โดยหาก y^* มีค่ามากกว่า τ จะได้ค่าสังเกต y=1 (แสดงโดยพื้นที่แรงเงา) และหาก y^* มีค่าน้อยกว่า τ ค่าสังเกตได้จะเท่ากับศูนย์ (y=0) และเนื่องจาก y^* เป็นค่าที่ไม่สามารถสังเกตได้ เราจึงไม่สามารถประมาณค่าความแปรปรวนของค่าคลาดเคลื่อน (variance of error) เหมือนในแบบจำลองสมการ

ถดถอยโดยทั่วไปได้ จำเป็นต้องกำหนดให้ความแปรปรวนเป็นค่าคงที่ค่าหนึ่ง ซึ่งในแบบจำลองโพรบิทกำหนดให้ $\text{var}(\varepsilon|X)=1$ ส่วนในแบบจำลองโลจิตกำหนดให้ $\text{var}(\varepsilon|X)=\pi^2/3\approx 3.29$

ภาพที่ ก.2
การแจกแจงของ y^* ในแบบจำลองตัวแปรตามที่มีลักษณะ
ข้อมูลแบบสองทางเลือก



ที่มา : Long (1997) หน้า 41

ทั้งนี้ ความน่าจะเป็นที่ y จะเท่ากับ 1 คือ

$$Prob\left(y=1|X_i\right)=Prob\left(y^*>0|X_i\right)=Prob\left(\varepsilon>-X_i'\beta|X_i\right)$$

เนื่องจาก การแจกแจงแบบปกติมีการกระจายข้อมูลแบบสมมาตร (symmetric) ทำให้

$$Prob\left(y=1|X_i\right)=Prob\left(y^*>0|X_i\right)=Prob\left(\varepsilon<X_i'\beta|X_i\right)=\Phi\left(X_i'\beta\right) \quad \text{(ก.15)}$$

โดยที่ $Prob\left(y=1\middle|X_i\right)$ คือ ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้น เมื่อกำหนดค่า X_i มาให้ และ $\Phi\left(X_i'\beta\right)$คือ ฟังก์ชันการแจกแจงค่าน่าจะเป็นสะสมแบบมาตรฐาน (standard normal CDF) แสดงโดย

$$\begin{aligned} Prob\left(y=1\middle|X_i\right) &= \int_{-\infty}^{X_i'\beta} \phi(t)dt = \Phi\left(X_i'\beta\right) \\ &= \int_{-\infty}^{X_i'\beta} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-(t-\mu)^2/2\sigma^2} dt \end{aligned} \quad (ก.16)$$

เนื่องจาก $\varepsilon \sim N\left(0,1\right)$ แทนค่า $\mu=0,\ \sigma^2=1$ จะได้

$$\begin{aligned} Prob\left(y=1\middle|X_i\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_i'\beta} e^{-t^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \left(-\frac{e^{-t^2/2}}{t} \right)\Bigg|_{t=-\infty}^{X_i'\beta} \\ &= -\frac{1}{\sqrt{2\pi}} \left[-\frac{e^{-(X_i'\beta)^2/2}}{X_i'\beta} + \frac{1}{-\infty e^{\infty}} \right] \\ &= -\frac{1}{\sqrt{2\pi}} \left(-\frac{e^{-(X_i'\beta)^2/2}}{X_i'\beta} \right) \end{aligned}$$

จะเห็นได้ว่าค่าความน่าจะเป็นมีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X ซึ่งสอดคล้องกับเงื่อนไขของแบบจำลองแบบสองทางเลือก และเนื่องจากแบบจำลองมีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X และพารามิเตอร์ ทำให้ไม่สามารถใช้ขั้นตอนการประมาณค่าแบบกำลังสองน้อยที่สุด (OLS) ได้ ซึ่งในกรณีเช่นนี้ต้องอาศัยวิธี maximum likelihood ในการประมาณค่าพารามิเตอร์ของแบบจำลอง โดยสมมติว่ามีตัวแปรสุ่ม n ตัวอย่าง ซึ่งแต่ละค่าที่สังเกตได้มีการแจกแจงแบบ Bernoulli และให้ $F(X_i'\beta)$ คือ ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้น ซึ่งมี joint probability density function หรือ likelihood function คือ

$$Prob(Y_1=y_1, Y_2=y_2, \ldots, Y_n=y_n \middle| X_i) = L = \prod_{y_i=0} \left(1-F\left(X_i'\beta\right)\right) \prod_{y_i=1} F\left(X_i'\beta\right) \quad (ก.17)$$

สำหรับแบบจำลองโพรบิท ซึ่งมีการแจกแจงแบบปกติจะมี likelihood function คือ

$$L=\prod_{y_i=0}\left(1-\Phi\left(X_i'\beta\right)\right)\prod_{y_i=1}\Phi\left(X_i'\beta\right)$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$\ln L=\sum_{y_i=0}\ln\left(1-\Phi\left(X_i'\beta\right)\right)+\sum_{y_i=1}\ln\Phi\left(X_i'\beta\right) \qquad \text{(ก.18)}$$

จะเห็นได้ว่า log likelihood function เป็นฟังก์ชันของพารามิเตอร์ β เนื่องจากค่า X_i เป็นค่าที่ทราบอยู่แล้ว ภายใต้เงื่อนไขลำดับที่หนึ่ง (first order condition) ของสมการที่ ก.18 เพื่อหาค่าพารามิเตอร์ที่ไม่ทราบค่า ที่ทำให้ log likelihood function มีสูงสุด จะได้ว่า

$$\frac{\partial \ln L}{\partial\beta}=\sum_{y_i=0}\frac{-\phi\left(X_i'\beta\right)}{1-\Phi\left(X_i'\beta\right)}X_i+\sum_{y_i=1}\frac{\phi\left(X_i'\beta\right)}{\Phi\left(X_i'\beta\right)}X_i=0 \qquad \text{(ก.19)}$$

ทำการแก้สมการ ก.19 เพื่อหาค่าประมาณ $\hat{\beta}$ ทั้งนี้ ในขั้นตอนการคำนวณจะมีลักษณะไม่เป็นเชิงเส้นกับพารามิเตอร์สูง (highly nonlinear in parameter) ทำให้จำเป็นต้องใช้วิธีการประมาณค่าสมการแบบไม่เป็นเชิงเส้นเพื่อให้ได้คำตอบที่เป็นตัวเลข ซึ่งเมื่อแทนค่าพารามิเตอร์ไม่ทราบค่าที่ได้จากการประมาณจะได้สมการประมาณค่าแบบจำลองโพรบิท ดังนี้

$$Prob\left(y=1\middle|X_i\right)=\Phi\left(X_i'\hat{\beta}\right) \qquad \text{(ก.20)}$$

ในการแปลความหมายแบบจำลองโพรบิทจะอธิบายโดยใช้ค่าผลกระทบส่วนเพิ่ม (marginal effect) แสดงโดย

$$E\left(y\middle|X_i\right)=F\left(X_i'\beta\right)$$

$$\frac{\partial E\left(y\middle|X_i\right)}{dX_i}=\frac{\partial F\left(X_i'\beta\right)}{\partial\left(X_i'\beta\right)}\cdot\frac{d\left(X_i'\beta\right)}{dX_i}=f\left(X_i'\beta\right)\beta$$

โดยที่ f(.) คือ density function ของฟังก์ชันการแจกแจงความน่าจะเป็นสะสม ,F(.) สำหรับการแจกแจงแบบปกติ ค่าผลกระทบส่วนเพิ่ม (marginal effect) คือ

$$\frac{\partial E\left(y|X_i\right)}{dX_i} = \phi\left(X_i'\beta\right)\beta \qquad \text{(ก.21)}$$

โดยที่ $\phi\left(X_i'\beta\right)$ คือ standard normal density function สมการที่ ก.21 แสดงถึงค่าผลกระทบส่วนเพิ่ม (marginal effect) ของแบบจำลองโพรบิท ซึ่งจะเห็นว่า อัตราการเปลี่ยนแปลงของความน่าจะเป็นเนื่องจากการเปลี่ยนแปลงในค่า X ประกอบด้วย ค่า β และฟังก์ชัน $\phi\left(X_i'\beta\right)$ ซึ่งเปลี่ยนแปลงตามค่า X_i ทำให้อัตราการเปลี่ยนแปลงของความน่าจะเป็นที่ได้มีลักษณะไม่เป็นเชิงเส้นกับตัวแปร X ทั้งนี้ หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงของค่า X 1 หน่วย จะส่งผลให้ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นเพิ่มขึ้น $\phi\left(X_i'\beta\right)\beta \times 100\%$ สำหรับการประเมินความน่าเชื่อถือของแบบจำลอง (goodness of fit) แบบจำลองโพรบิทมักนิยมใช้ค่า pseudo R^2 แทนค่า R^2 เช่นเดียวกับแบบจำลองโลจิต

<u>เปรียบเทียบผลกระทบส่วนเพิ่ม (marginal effect) ของแบบจำลองแบบสองทางเลือก</u>

ในแบบจำลองสมการถดถอยเชิงเส้น (linear regression model) ค่าสัมประสิทธิ์ความชัน (slope coefficient) จะแสดงถึง การเปลี่ยนแปลงในค่าเฉลี่ยของตัวแปรตามอันเนื่องมาจากการเปลี่ยนแปลงในค่าของตัวแปรอธิบาย 1 หน่วย โดยกำหนดให้ปัจจัยอื่นๆคงที่ สำหรับแบบจำลองที่ตัวแปรตามมีลักษณะข้อมูลแบบสองทางเลือก (dichotomous dependent variable) ซึ่งที่ผ่านมาได้กล่าวถึงแนวคิดของแบบจำลองที่เป็นที่นิยม 3 แนวคิด ได้แก่ แบบจำลองความน่าจะเป็นเชิงเส้น (linear probability model), แบบจำลองโลจิต (logit model) และแบบจำลองโพรบิท (probit model) โดยในแบบจำลองความน่าจะเป็นเชิงเส้น ค่าสัมประสิทธิ์ความชันที่คำนวณได้จะส่งผลโดยตรงต่อความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้น หากค่าของตัวแปรอธิบายเปลี่ยนแปลงไป 1 หน่วย โดยกำหนดให้ผลของปัจจัยอื่นๆคงที่ เช่น หาก $\beta_2 = 0.1$, y คือ การตัดสินใจทำการวิจัยและพัฒนาของบริษัท และ X คือ ขนาดของบริษัท โดยวัดจากจำนวนพนักงาน จะได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การที่บริษัทมีพนักงานเพิ่มขึ้น 1 คน จะส่งผลให้ความน่าจะเป็นที่บริษัทจะตัดสินใจทำการวิจัยและพัฒนาเพิ่มขึ้น 0.1

สำหรับแบบจำลองโลจิต ค่าสัมประสิทธิ์ความชันที่คำนวณได้ แสดงถึงการเปลี่ยนแปลงของ log of odds ratio เนื่องมาจากการเปลี่ยนแปลงในค่าของตัวแปรอธิบาย 1

หน่วย ซึ่งหากทำการ antilog ค่าที่คำนวณได้ จะได้ว่า การเปลี่ยนแปลงเพิ่มขึ้นของค่า X 1 หน่วย จะส่งผลให้อัตราส่วนความเป็นไปได้ (odds ratio) เพิ่มขึ้น $e^{\hat{\beta}_2}$ เท่า หรือ $\left(e^{\hat{\beta}_2}-1\right)\times 100\%$ นอกจากนี้ ยังสามารถคำนวณหาอัตราการเปลี่ยนแปลงของความน่าจะเป็น (rate of change of probability) ที่เหตุการณ์จะเกิดขึ้น หรือค่าผลกระทบส่วนเพิ่ม (marginal effect) แสดงโดย

$$\frac{dE\left(y|X_i\right)}{dX}=\frac{dP_i}{dX}=\beta_j P_i\left(1-P_i\right)$$

โดยที่ β_j คือ ค่าสัมประสิทธิ์ของตัวแปรอธิบายที่ j ซึ่งในการคำนวณค่า P_i จะคำนวณจากตัวแปรตามทุกตัวที่ใช้ในการวิเคราะห์ โดย $\hat{P}_i=\frac{1}{1+e^{-(X\hat{\beta})}}$ สำหรับการแปลความหมายอธิบายได้ว่า หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงเพิ่มขึ้นของค่า X 1 หน่วยจะส่งผลให้ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นเพิ่มขึ้น $\hat{\beta}_j\hat{P}_i\left(1-\hat{P}_i\right)\times 100\%$

ส่วนในแบบจำลองโพรบิท ในการแปลความหมายจะอธิบายโดยใช้ค่าผลกระทบส่วนเพิ่ม (marginal effect) ซึ่งค่อนข้างจะมีความซับซ้อน โดยที่

$$\frac{dP_i}{dX_i}=\frac{\partial E\left(y|X_i\right)}{dX}=\phi\left(X_i'\beta\right)\beta$$

โดยที่ $\phi\left(X_i'\beta\right)$ คือฟังก์ชันความหนาแน่นแบบมาตรฐาน (standard normal density function) และ $X'\beta=\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki}$ ส่วนการแปลความหมายจะอธิบายเหมือนกับแบบจำลองโลจิต คือ หากกำหนดให้ปัจจัยอื่นคงที่ การเปลี่ยนแปลงเพิ่มขึ้นของค่า X 1 หน่วยจะส่งผลให้ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นเพิ่มขึ้น $\phi\left(X_i'\hat{\beta}\right)\hat{\beta}\times 100\%$ ทั้งนี้ จะเห็นได้ว่า การคำนวณค่าผลกระทบส่วนเพิ่มของทั้งแบบจำลองโลจิตและโพรบิทต่างคำนวณจากตัวแปรอธิบายทุกตัวที่ใช้ในแบบจำลอง ขณะที่แบบจำลองความน่าจะเป็นแบบเชิงเส้นจะคำนวณค่าผลกระทบส่วนเพิ่มจากตัวแปรอธิบายที่สนใจเพียงตัวเดียวเท่านั้น

เนื่องจาก ในการคำนวณค่าผลกระทบส่วนเพิ่มของแบบจำลองโพรบิทและโลจิตจำเป็นต้องอาศัยค่าของตัวแปรอธิบายทุกตัวที่ใช้ในแบบจำลอง ซึ่งในการเลือกค่าของตัวแปรอธิบายเพื่อใช้ในการคำนวณผลกระทบส่วนเพิ่มจะเลือกใช้ค่าเฉลี่ยของตัวอย่าง (sample mean) ทั้งนี้ ในการคำนวณค่าผลกระทบส่วนเพิ่มในกรณีที่ตัวแปรอธิบายเป็นตัวแปรหุ่น (dummy

variable) จะมีการคำนวณที่แตกต่างออกไป โดยหากกำหนดให้ x_d คือ ตัวแปรหุ่น, $\bar{x}_d$ คือ สัดส่วนของตัวอย่างที่ $x_d=1$ ผลกระทบส่วนเพิ่มในกรณีที่ตัวแปรอธิบายเป็นตัวแปรหุ่น จะคำนวณจากผลต่างระหว่างความน่าเป็นที่เหตุการณ์จะเกิดขึ้นกับความน่าเป็นที่เหตุการณ์จะไม่เกิดขึ้น แสดงโดย

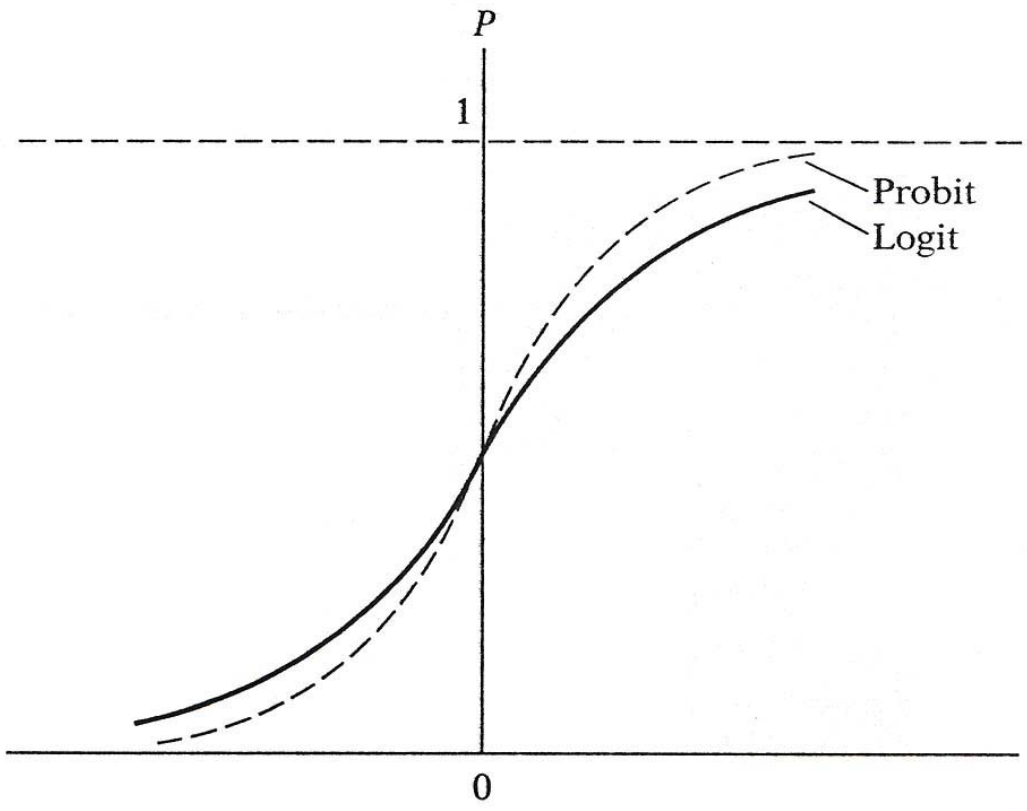
$$marginal\ effect = Prob\left(y=1|\bar{X}, x_d=1\right) - Prob\left(y=1|\bar{X}, x_d=0\right)$$

สำหรับการแปลความหมายอธิบายได้ว่า หากกำหนดให้ตัวแปรอื่นคงที่ เมื่อ $x_d=1$ จะส่งผลให้ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นมีมากกว่า กรณีที่ $x_d=0$ เท่ากับ $\Delta Prob\left(y=1|\bar{X}\right)$ เช่น สมมติว่าตัวแปรหุ่น คือ การได้รับเงินอุดหนุนจากภาครัฐ และตัวแปรตามคือ ความน่าจะเป็นที่หน่วยผลิตจะตัดสินใจทำการวิจัยและพัฒนา จะได้ว่า หากกำหนดให้ตัวแปรอื่นคงที่ บริษัทที่ได้รับเงินอุดหนุนจากภาครัฐมีโอกาสที่จะทำการวิจัยและพัฒนามากกว่าบริษัทที่ไม่ได้รับเงินอุดหนุนจากภาครัฐ เท่ากับ $\Delta Prob\left(y=1|\bar{X}\right)$

อย่างไรก็ตาม ถึงแม้ว่าแบบจำลองโพรบิทและโลจิตจะถูกนำมาปรับประยุกต์ใช้และให้ผลการประมาณค่าที่ค่อนข้างคล้ายคลึงกัน แต่ก็มีความแตกต่างกันในส่วนของลักษณะการกระจายตัวของฟังก์ชันการแจกแจงความน่าจะเป็นสะสม (CDF) ของแบบจำลองโพรบิทและโลจิต โดยปลายของเส้นการแจกแจงแบบโลจิตจะมีความหนากว่าเส้นการแจกแจงแบบโพรบิทเล็กน้อย เนื่องจากแบบจำลองโลจิตมีความแปรปรวนมากกว่าแบบจำลองโพรบิท ดังแสดงในภาพที่ ก.3 ซึ่งจะเห็นว่า เมื่อค่าความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) มีค่าเข้าใกล้ 0 หรือ 1 แบบจำลองโลจิตจะเข้าใกล้ค่าดังกล่าวด้วยอัตราที่ต่ำกว่าแบบจำลองโพรบิท ดังแสดงในตาราง ก.1 อย่างไรก็ตาม เราไม่สามารถเปรียบเทียบค่าสัมประสิทธิ์, β, ของแบบจำลองทั้งสองโดยตรงได้ เนื่องจากแบบจำลองทั้งสองมีค่าความแปรปรวนของการแจกแจง (variance of distribution) ที่ต่างกัน โดยการแจกแจงแบบปกติมีค่าความแปรปรวนเท่ากับ 1 ขณะที่การแจกแจงแบบโลจิสติกมีค่าความแปรปรวนเท่ากับ $\pi^2/3$ อย่างไรก็ตาม หากเราคูณค่าสัมประสิทธิ์ของแบบจำลองโพรบิทด้วยค่า $\pi/\sqrt{3} \approx 1.8$ จะทำให้เราได้ค่าโดยประมาณของค่าสัมประสิทธิ์ของแบบจำลองโลจิต ทั้งนี้ Amemiya (1981) แนะนำว่า หากคูณค่าสัมประสิทธิ์ของแบบจำลองโพรบิทด้วยค่า 1.6 จะทำให้เราได้ค่าโดยประมาณของค่าสัมประสิทธิ์ของแบบจำลองโลจิตที่ดีกว่า ดังนั้น จึงเป็นการยากที่จะเลือกใช้แบบจำลองใดแบบจำลองหนึ่งในการประมาณค่าแบบจำลองที่ตัวแปรตามมีลักษณะข้อมูลแบบสองทางเลือก ซึ่งในทางปฏิบัติแบบจำลองโลจิตจะมีรูปแบบการคำนวณทางคณิตศาสตร์ที่ง่ายกว่าแบบจำลองโพรบิท

ภาพที่ ก.3

เปรียบเทียบการแจกแจงสะสมแบบปกติและแบบโลจิสติก



ที่มา : Gujarati (2003) หน้า 614

ตารางที่ ก.1

ค่าของฟังชันการแจกแจงความน่าจะเป็นสะสมที่ระดับต่างๆ

	Cumulative normal	Cumulative logistic
Z	$P_1(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-t^2/2}dt$	$P_2(Z)=\frac{1}{1+e^{-Z}}$
-3.0	0.0013	0.0474
-2.0	0.0228	0.1192
-1.5	0.0668	0.1824
-1.0	0.1587	0.2689
-0.5	0.3085	0.3775
0.0	0.5000	0.5000
0.5	0.6915	0.6225
1.0	0.8413	0.7311
1.5	0.9332	0.8176
2.0	0.9772	0.8808
3.0	0.9987	0.9526

ที่มา : Gujarati (2003) หน้า 615

3. <u>แบบจำลองโทบิท (Tobit model)</u>

ตัวแปรตามที่มีค่าต่อเนื่องในบางครั้งจะมีค่าในช่วงปลายที่ขาดหายไป อาจเนื่องมาจากการไม่สามารถวัดหรือสังเกตค่าดังกล่าวได้ เราจึงพบว่ามีตัวแปรตามที่มีค่าเท่ากับศูนย์มีจำนวนมากพอสมควร เช่น ค่าใช้จ่ายในการวิจัยและพัฒนา ค่าใช้จ่ายในการซื้อบ้าน เป็นต้น ซึ่งแบบจำลองโทบิทเป็นแบบจำลองที่เหมาะสมสำหรับสถานการณ์ดังกล่าวนี้ โดยแบบจำลองนี้ถูกนำเสนอครั้งแรกโดย Tobin (1958) ซึ่งแบบจำลองโทบิทเป็นการต่อยอดแบบจำลองโพรบิท โดยในแบบจำลองโพรบิทเป็นการประมาณค่าความน่าจะเป็นที่บริษัทจะตัดสินใจทำการวิจัยและพัฒนา ขณะที่แบบจำลองโทบิทจะสนใจในจำนวนเงินที่บริษัทใช้ไปเพื่อการวิจัยและพัฒนา อย่างไรก็ตาม เราต้องเผชิญกับสถานการณ์ที่จะต้องเลือก (dilemma) ซึ่งหากบริษัทตัดสินใจไม่ทำ

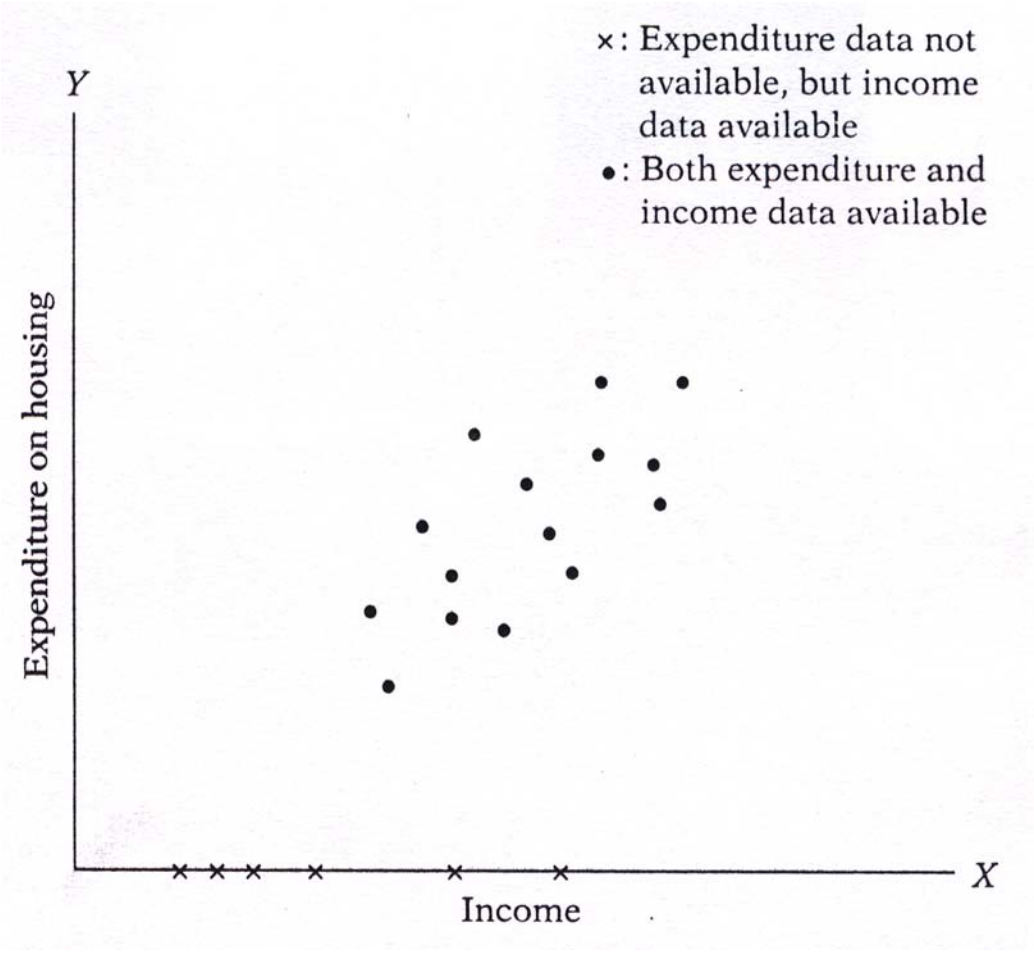
การวิจัยและพัฒนา ก็จะไม่มีข้อมูลค่าใช้จ่ายในการวิจัยและพัฒนาของบริษัท ซึ่งข้อมูลนี้จะมีเฉพาะในบริษัทที่ตัดสินใจทำการวิจัยและพัฒนาเท่านั้น

ทั้งนี้ สามารถแบ่งกลุ่มตัวอย่างออกเป็น 2 กลุ่ม โดยในกลุ่มแรก, n_1 , คือ กลุ่มที่มีทั้งข้อมูลตัวแปรตาม (ค่าใช้จ่ายในการวิจัยและพัฒนา) และตัวแปรอธิบาย (เช่น ขนาดของหน่วยผลิต, การส่งออกสินค้า เป็นต้น) ขณะที่กลุ่มที่สอง, n_2 , มีเพียงข้อมูลตัวแปรอธิบายเท่านั้น ไม่มีข้อมูลของตัวแปรตาม ซึ่งการที่ข้อมูลตัวแปรตามสามารถสังเกตได้เฉพาะบางข้อมูลเราเรียกว่าข้อมูลที่ถูกเซนเซอร์ (censored data) ดังนั้น แบบจำลองโทบิทจึงถูกเรียกว่า แบบจำลองสมการถดถอยที่ถูกเซนเซอร์ (censored regression model) หรือแบบจำลองถดถอยที่ตัวแปรตามถูกจำกัด (limited dependent variable regression model) ซึ่งเป็นเพราะการจำกัดค่าของตัวแปรตามที่สังเกตได้

ในการประมาณค่าแบบจำลองเราไม่สามารถเลือกเฉพาะกลุ่มตัวอย่าง n_1 มาใช้ในการประมาณค่าโดยไม่สนใจกลุ่มตัวอย่าง n_2 ได้ ซึ่งในการประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จะประมาณค่าพารามิเตอร์โดยอาศัยข้อมูลที่สังเกตได้จากกลุ่มตัวอย่าง n_1 ทำให้ค่าประมาณที่ได้มีความเอนเอียง (bias) และไม่น่าเชื่อถือ (inconsistent) โดยความเอนเอียงจะเกิดขึ้นหากเราพิจารณาเฉพาะข้อมูลจากกลุ่มตัวอย่าง n_1 และละเลยข้อมูลจากกลุ่มตัวอย่าง n_2 ทำให้ไม่สามารถรับประกันได้ว่า $E(u_i)$ จะยังคงมีค่าเท่ากับศูนย์ และการที่ $E(u_i) \neq 0$ ก็ไม่สามารถรับประกันได้ว่าการประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จะยังคงมีความไม่เอนเอียง (unbiased) ทั้งนี้ในการอธิบายสามารถดูได้จากภาพที่ ก.4 โดยที่กำหนดให้เครื่องหมายกากบาท คือ ค่าจากกลุ่มตัวอย่าง n_2 ที่ซึ่งตัวแปรตาม y ไม่สามารถสังเกตได้ (เนื่องจากถูกเซนเซอร์) โดยค่าดังกล่าวจะอยู่บนแกน X ทั้งหมด และหากตัวแปรตาม y สามารถสังเกตได้ จะนำข้อมูลจากกลุ่มตัวอย่าง n_1 แสดงโดยเครื่องหมายจุดซึ่งจะอยู่บนระนาบ X-Y ทั้งนี้จะเห็นได้ว่า หากทำการประมาณค่าเฉพาะข้อมูลจากกลุ่มตัวอย่าง n_1 ผลการประมาณค่าจุดตัดแกน (intercept) และค่าความชันของสัมประสิทธิ์ที่ได้จะมีความแตกต่างจากการใช้ข้อมูลทั้งหมด (n_1+ n_2) มาประมาณค่าเส้นถดถอย (regression line)

นอกจากนี้ หากพิจารณาการตัดปลายของการแจกแจงแบบปกติ (truncated normal distribution) เมื่อจุดตัดปลายเท่ากับ -0.5, 0 และ 0.5 ดังแสดงในภาพที่ ก.5 จะเห็นได้ว่า ค่าเฉลี่ย (μ) ของการแจกแจงแบบปกติจะไม่เท่ากับศูนย์เช่นเดิมอีกต่อไป โดยเมื่อจุดตัดปลายมีค่าเพิ่มขึ้นค่าเฉลี่ยก็จะมีค่ามากกว่าค่าเฉลี่ยที่จุดกำเนิด ขณะที่ความแปรปรวนกลับมีค่าลดลง แสดงโดยสมการทางคณิตศาสตร์ดังนี้

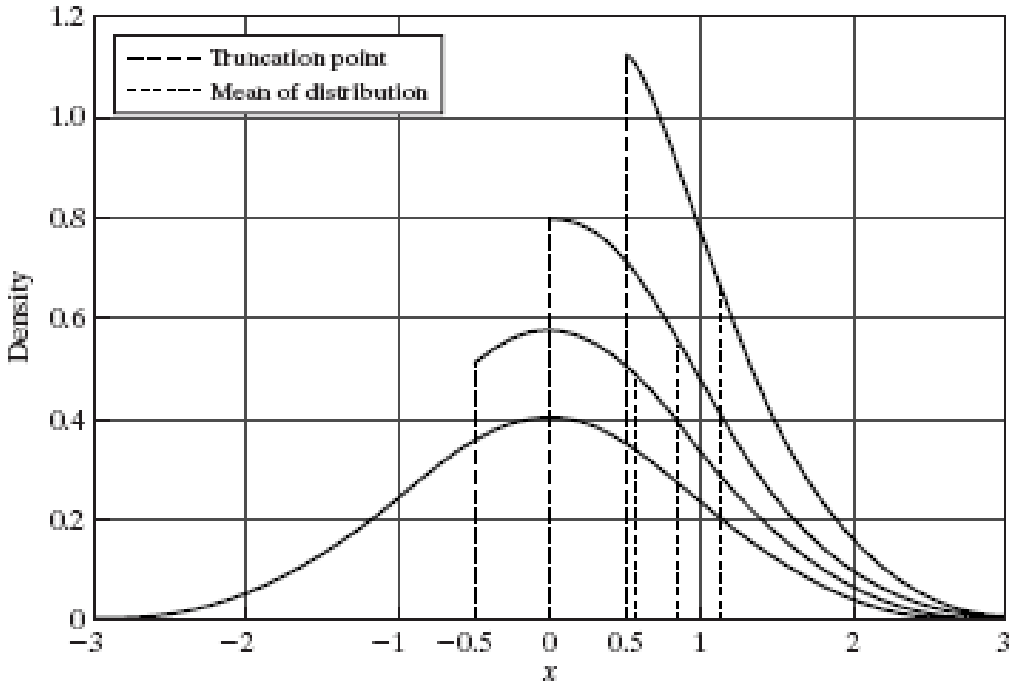
ภาพที่ ก.4
แผนภาพการลงข้อมูลที่สังเกตมาได้



ที่มา : Gujarati (2003) หน้า 617

ภาพที่ ก.5

การแจกแจงแบบปกติที่ตัดปลาย (Truncated normal distribution)



ที่มา : Greene (2002) หน้า 758

จากสมการถดถอยทั่วไป

$$y_i = X_i'\beta + \varepsilon_i$$

หาก $\varepsilon_i | X_i \sim N\left(0, \sigma^2\right)$ แล้วจะได้ว่า $y | X_i \sim N\left(X_i'\beta, \sigma^2\right)$ เนื่องจากเราสนใจในการแจกแจงของ y_i เมื่อ y_i มีค่ามากกว่าค่าคงที่ a ซึ่งจากทฤษฎีโมเมนต์การตัดปลายของการแจกแจงแบบปกติ (moment of truncated normal distribution)[2] จะได้ว่า ค่าเฉลี่ยและความแปรปรวนของการตัดปลาย (truncation) มีค่าดังนี้

$$\begin{aligned} E\left(y_i | y_i > a\right) &= \mu + \sigma\lambda\left(\alpha_i\right) \\ &= X'\beta + \sigma\lambda\left(\alpha_i\right) \end{aligned} \qquad \text{(ก.22)}$$

[2] อ้างใน Greene (2002) หน้า 759

และมีความแปรปรวนเท่ากับ

$$\text{var}\left(y_i \middle| y_i > a\right) = \sigma^2\left[1-\delta\left(\alpha_i\right)\right] \qquad \text{(ก.23)}$$

โดยที่ $\alpha_i = \left(a-\mu\right)/\sigma$, $\phi\left(\alpha_i\right)$ คือ standard normal density และ

$$\lambda\left(\alpha_i\right) = \phi\left(\alpha_i\right)/\left[1-\Phi\left(\alpha_i\right)\right]$$
$$\delta\left(\alpha_i\right) = \lambda\left(\alpha_i\right)\left[\lambda\left(\alpha_i\right)-\alpha_i\right]$$

ทั้งนี้ค่า $\delta\left(\alpha_i\right)$ มีค่าอยู่ระหว่าง 0 ถึง 1 , $0 < \delta\left(\alpha_i\right) < 1$, และเรียก $\lambda\left(\alpha_i\right)$ ว่า inverse Mills ratio เราสามารถเขียนสมการถดถอยสำหรับการตัดปลายได้ดังนี้

$$y_i \middle| y_i > a = E\left(y_i \middle| y_i > a\right) + u_i = X'\beta + \sigma\lambda\left(\alpha_i\right) \qquad \text{(ก.24)}$$

ทั้งนี้ u_i จะยังคงมีค่าเฉลี่ยเท่ากับศูนย์แต่จะมีความแปรปรวนไม่คงที่ (heteroscedastic) ดังนี้

$$\text{var}\left(u_i\right) = \sigma^2\left[1-\delta\left(\alpha_i\right)\right]$$

จะเห็นได้ว่า หากประมาณค่าสมการ ก.24 ด้วยวิธีกำลังสองน้อยที่สุด (OLS) จะทำให้เราเพิกเฉยต่อตัวแปรไม่เป็นเชิงเส้น $\lambda\left(\alpha_i\right)$ เป็นผลให้เกิดการเอนเอียง (bias) เนื่องมาจากการละทิ้งตัวแปร (omitted variable) ซึ่งจากปัญหาที่กล่าวมาในเรื่องของความเอนเอียงและความไม่น่าเชื่อถือของการประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จึงเกิดแนวคิดในการพัฒนาแบบจำลองขึ้นมาใหม่ เรียกว่าแบบจำลองโทบิท ซึ่งเป็นแบบจำลองถดถอยสำหรับตัวแปรตามที่ถูกจำกัด โดยอาศัยวิธีการประมาณค่าแบบ maximum likelihood โดยมีหลักการดังนี้

สมมติว่ามีตัวแปรแฝง (latent variable) ที่ไม่สามารถสังเกตได้ , y_i^* , ซึ่งมีความสัมพันธ์เชิงเส้นและขึ้นอยู่กับเวกเตอร์ตัวแปรอธิบาย X_i ผ่านทางพารามิเตอร์ β ซึ่งเป็นตัวกำหนดความสัมพันธ์ระหว่างตัวแปรอธิบาย , X_i , กับตัวแปรแฝง , y_i^* , หากกำหนดให้ค่าความคลาดเคลื่อน (error term) , u_i , มีการแจกแจงแบบปกติ และนิยามตัวแปรที่สังเกตได้ y_i มีค่าเท่ากับตัวแปรแฝง หากตัวแปรแฝงมีค่ามากกว่าศูนย์ และเท่ากับศูนย์หากตัวแปรแฝงมีค่าน้อยกว่าหรือเท่ากับศูนย์ ดังนี้

$$y_i = \begin{cases} y_i^* \ \ if \ \ y_i^* > 0 \\ 0 \ \ if \ \ y_i^* \le 0 \end{cases} \qquad \text{(ก.25)}$$

โดยที่ $y_i^* = X_i'\beta + u_i$

และ $u_i \sim N\left(0, \sigma^2\right)$, $y_i^* \sim N\left(X_i'\beta, \sigma^2\right)$

เนื่องจากข้อมูลตัวแปรตามที่สังเกตได้ถูกเซนเซอร์ทำให้การแจกแจงของแบบจำลองเป็นการผสมกันระหว่างการแจกแจงแบบไม่ต่อเนื่อง (discrete distribution) และการแจกแจงแบบต่อเนื่อง (continuous distribution) ซึ่งหาก y_i^* มีการแจกแจงแบบปกติแล้ว ความน่าจะเป็นที่ค่าที่สังเกตได้จะถูกเซนเซอร์เท่ากับ

$$Prob\left(Censored\right) = Prob\left(y^* \le 0\right) = \Phi(-X'\beta/\sigma) = 1 - \Phi\left(X'\beta/\sigma\right)$$

และความน่าจะเป็นที่ค่าที่สังเกตได้จะไม่ถูกเซนเซอร์เท่ากับ

$$Prob\left(Uncensored\right) = Prob\left(y^* > 0\right) = 1 - \Phi\left(-X'\beta/\sigma\right) = \Phi(X'\beta/\sigma)$$

ทั้งนี้ ค่าคาดหวังแบบมีเงื่อนไขของ y_i เมื่อกำหนด X_i เท่ากับ

$$\begin{aligned} E\left(y_i \middle| X_i\right) &= Prob\left(y_i \le 0\right) \times E\left(y_i \middle| X_i, y_i = 0\right) + Prob\left(y_i > 0\right) \times E\left(y_i \middle| X_i, y_i > 0\right) \\ &= Prob\left(y_i \le 0\right) \times 0 + Prob\left(y_i^* > 0\right) \times E\left(y_i^* \middle| X_i, y_i^* > 0\right) \\ &= \Phi\left(\frac{X_i'\beta}{\sigma}\right)\left(X_i'\beta + \sigma\lambda\left(\alpha_i\right)\right) \end{aligned} \qquad \text{(ก.26)}$$

โดยที่ $\lambda\left(\alpha_i\right) = \dfrac{\phi\left[\left(0 - X'\beta_i\right)/\sigma\right]}{1 - \Phi\left[\left(0 - X'\beta_i\right)/\sigma\right]} = \dfrac{\phi\left(X'\beta_i/\sigma\right)}{\Phi\left(X'\beta_i/\sigma\right)}$

และมีค่าความแปรปรวนคือ

$$\text{var}\left(y_i \middle| X_i\right) = \sigma^2\left(1 - \Phi\left(\alpha_i\right)\right)\left[1 - \delta\left(\alpha_i\right) + \left(\alpha_i + \lambda\left(\alpha_i\right)\right)^2 \Phi\left(\alpha_i\right)\right]$$

สำหรับการประมาณค่าแบบจำลองโทบิทอาศัยวิธีการประมาณค่าแบบ maximum likelihood ซึ่งจากการที่แบบจำลองโทบิทประกอบด้วยกลุ่มตัวอย่าง 2 กลุ่ม คือ กลุ่มตัวอย่างข้อมูลที่ไม่ถูกเซนเซอร์ ซึ่งจะมีรูปแบบของการประมาณค่าเช่นเดียวกับในแบบจำลองเชิงเส้นทั่วไป ขณะที่กลุ่มตัวอย่างที่สอง คือ กลุ่มตัวอย่างข้อมูลที่ถูกเซนเซอร์ ซึ่งกลุ่มตัวอย่างนี้จะไม่สามารถสังเกตค่า y* ที่แท้จริงได้ ทราบเพียงแต่ว่ามีค่าน้อยกว่าศูนย์, $y^* \le 0$ ดังนั้น เราจึงใช้ความน่าจะเป็นที่ข้อมูลจะถูกเซนเซอร์ที่คำนวณได้แทนในสมการ likelihood ซึ่งจากสมการถดถอยสำหรับข้อมูลที่ไม่ถูกเซนเซอร์

$$y_i = X_i'\beta + u_i \qquad \text{เมื่อ } y_i^* > 0$$

โดยที่ $u_i \sim N\left(0, \sigma^2\right)$, $y_i^* \sim N\left(X_i'\beta, \sigma^2\right)$ จะได้ว่า pdf ของ y_i คือ

$$\begin{aligned} f(y_i \mid X_i'\beta, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\cdot\frac{(y_i - X_i'\beta)^2}{\sigma^2}} \\ &= \frac{1}{\sigma}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - X_i'\beta}{\sigma}\right)^2}\right] \\ &= \frac{1}{\sigma}\phi\left(\frac{y_i - X_i'\beta}{\sigma}\right) \end{aligned}$$

ทั้งนี้ สมการ likelihood สำหรับกลุ่มตัวอย่างที่ไม่ถูกเซนเซอร์สามารถเขียนได้ ดังนี้

$$L_U(\beta, \sigma^2) = \prod_{uncensored} \frac{1}{\sigma}\phi\left(\frac{y_i - X_i'\beta}{\sigma}\right) \qquad \text{(ก.27)}$$

เมื่อแปลงให้อยู่ในรูปลอการิทึมจะได้ว่า

$$\ln L_U(\beta, \sigma^2) = \sum_{uncensored} \ln \frac{1}{\sigma}\phi\left(\frac{y_i - X_i'\beta}{\sigma}\right)$$

สำหรับกลุ่มตัวอย่างที่ถูกเซนเซอร์ เราจะสนใจค่าของพื้นที่ใต้กราฟเมื่อ $y \le 0$ มากกว่า pdf ของ y เราจึงใช้ค่าความน่าจะเป็นที่ข้อมูลจะถูกเซนเซอร์ในสมการ likelihood แทน ดังนี้

$$L_C(\beta,\sigma^2) = \prod_{censored} 1-\Phi\left(\frac{X_i'\beta}{\sigma}\right) \qquad \text{(ก.28)}$$

เมื่อแปลงให้อยู่ในรูปลอการิทึมจะได้ว่า

$$\ln L_C(\beta,\sigma^2) = \sum_{censored} \ln\left[1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)\right]$$

รวมสมการทั้งสองเข้าด้วยกันจะได้ likelihood function สำหรับแบบจำลองโทบิท คือ

$$\begin{aligned} L(\beta,\sigma^2) &= \prod_{uncensored} \frac{1}{\sigma}\phi\left(\frac{y_i - X_i'\beta}{\sigma}\right)\cdot \prod_{censored} 1-\Phi\left(\frac{X_i'\beta}{\sigma}\right) \\ &= \prod_{y_i>0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - X'\beta)^2}{2\sigma^2}} \cdot \prod_{y_i=0}\left(1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)\right) \end{aligned} \qquad \text{(ก.29)}$$

เมื่อ $\Phi(.)$ คือ ฟังก์ชันการแจงแจงความน่าจะเป็นสะสมแบบมาตรฐาน (standard normal cumulative distribution function) และ $\phi(.)$ คือ ฟังก์ชันความหนาแน่นแบบมาตรฐาน (standard normal pdf) จะเห็นได้ว่า likelihood function ประกอบด้วยกัน 2 ส่วน คือ pdf ของกลุ่มตัวอย่างข้อมูลที่ไม่ถูกเซนเซอร์ (pdf for uncensored observation) และ cdf ของกลุ่มตัวอย่างข้อมูลที่ถูกเซนเซอร์ (cdf for censored observation) โดยส่วนแรกสอดคล้องกับสมการถดถอยสำหรับค่าสังเกตที่ไม่ถูกจำกัด (classical regression for non-limit observation) สำหรับส่วนที่สองจะเกี่ยวข้องกับความน่าจะเป็นสำหรับค่าสังเกตที่ถูกจำกัด (probability for limit observation) ทั้งนี้ หากแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\ln\left(2\pi\sigma^2\right)\right] - \frac{\left(y_i - X'\beta\right)^2}{2\sigma^2} + \sum_{y_i=0}\ln\left(1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)\right) \qquad \text{(ก.30)}$$

ภายใต้เงื่อนไขลำดับที่หนึ่ง (first order condition) ของสมการที่ ก.30 เพื่อหาค่าพารามิเตอร์ที่ไม่ทราบค่า ที่ทำให้ log likelihood function มีสูงสุด จะได้ว่า

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{y_i>0}\left(y_i - X_i'\beta\right)X_i - \sum_{y_i=0}\frac{1}{1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i'\beta)^2}{2\sigma^2}}\cdot\frac{X'}{\sigma^2} = 0 \qquad \text{(ก.31)}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_{y_i>0} -\frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{\left(y_i - X_i'\beta\right)^2}{\sigma^4}\right] + \sum_{y_i=0}\frac{1}{1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i'\beta)^2}{2\sigma^2}}\cdot\frac{\left(X_i'\beta\right)^2}{2\sigma^4} = 0 \quad \text{(ก.32)}$$

แก้สมการ ก.31 และ ก.32 เพื่อหาค่า $\hat{\beta}$ และ $\hat{\sigma}^2$ ซึ่งเป็นตัวประมาณของ β และ σ^2 ตามลำดับ จะได้สมการประมาณค่าแบบจำลองโทบิท คือ

$$\hat{y}_i^* = X_i'\hat{\beta} \qquad \text{(ก.33)}$$

ในการแปลความหมายแบบจำลองโทบิทจะอธิบายโดยใช้ค่าผลกระทบส่วนเพิ่ม (marginal effect) ซึ่งค่าที่ได้จะมีความแตกต่างจากค่าทั่วไป ทั้งนี้เนื่องจาก y_i^* เป็นค่าที่ไม่สามารถสังเกตได้ ทำให้ $\frac{\partial E\left(y_i^* \middle| X_i\right)}{dX_i} = \beta$ เป็นค่าที่ไม่ได้รับความสนใจ ซึ่งในการคำนวณค่าผลกระทบส่วนเพิ่มของแบบจำลองสมการถดถอยที่ถูกเซนเซอร์จะคำนวณจากค่าที่สังเกตได้ , y , แสดงโดย (เพื่อความสะดวก กำหนดให้ $\phi_i = \phi\left(\frac{X_i'\beta}{\sigma}\right)$ และ $\Phi_i = \Phi\left(\frac{X_i'\beta}{\sigma}\right)$

$$\frac{\partial E\left(y_i \middle| X_i\right)}{dX_i} = \frac{\partial\left[\Phi_i(X_i'\beta + \sigma\lambda(\alpha))\right]}{dX_i}$$

$$= \frac{\partial\Phi_i(X_i'\beta)}{dX_i} + \frac{\partial\left(\Phi_i\sigma\frac{\phi_i}{\Phi_i}\right)}{dX_i}$$

$$= \frac{\partial\Phi_i(X_i'\beta)}{dX_i} + \frac{\partial\sigma\phi_i}{dX_i}$$

$$= \Phi_i\beta + X_i'\beta\cdot\phi_i\cdot\frac{\beta}{\sigma} - \sigma\left(\frac{X_i'\beta}{\sigma}\cdot\phi_i\cdot\frac{\beta}{\sigma}\right)$$

$$= \Phi_i\beta + X_i'\beta\cdot\phi_i\cdot\frac{\beta}{\sigma} - X_i'\beta\cdot\phi_i\cdot\frac{\beta}{\sigma}$$

$$\frac{\partial E\left(y_i \middle| X_i\right)}{dX_i} = \Phi\left(\frac{X_i'\beta}{\sigma}\right)\beta = \Pr(Uncensored \middle| X)\cdot\beta \qquad \text{(ก.34)}$$

สมการที่ ก.34 แสดงให้เห็นว่า ค่าผลกระทบส่วนเพิ่ม (marginal effect) สามารถคำนวณได้จากค่าประมาณสัมประสิทธิ์ที่คำนวณได้คูณกับความน่าจะเป็นของค่าที่สังเกตได้จากกลุ่มตัวอย่างที่ไม่ถูกเซนเซอร์ (proportion of non-limit observation in sample) โดยในการแปลความหมายอธิบายได้ว่า หากกำหนดให้ตัวแปรอื่นคงที่ การเปลี่ยนแปลงในค่า X_i 1 หน่วย จะส่งผลให้ค่าของค่าคาดหวังแบบมีเงื่อนไข (conditional expected) ของตัวแปรตามเปลี่ยนแปลง $\Phi\left(\frac{X_i'\hat{\beta}}{\sigma}\right)\hat{\beta}$ หน่วย

4. <u>แบบจำลองทางเลือก (sample selection model)</u>

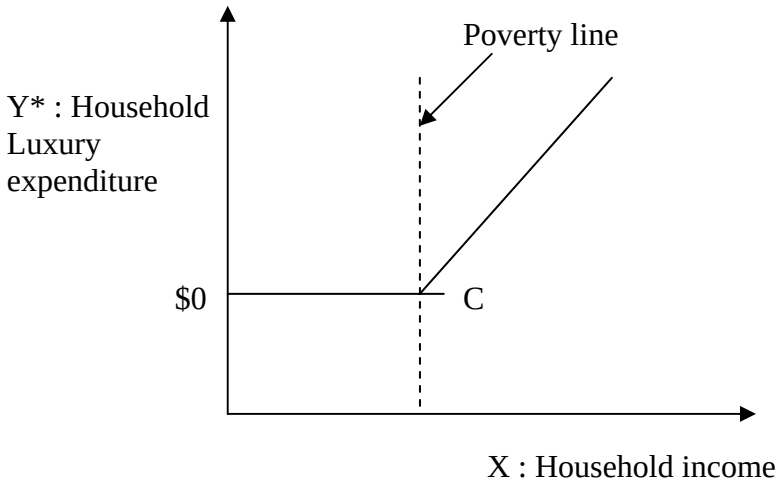
ในการวิเคราะห์แบบจำลองที่ตัวอย่างถูกเซนเซอร์ นอกจากจะวิเคราะห์โดยใช้แบบจำลองโทบิทแล้ว ยังสามารถประมาณค่าแบบจำลองทางเลือกอื่นที่อาศัยวิธี maximum likelihood ได้อีกด้วย โดย Heckman (1979) ได้เสนอแบบจำลองทางเลือก (sample selection model) หรือแบบจำลอง Heckman selection model ที่อธิบายถึง ความเอนเอียง (bias) ที่เกิดจากการเลือกตัวอย่าง (nonrandomly select sample) มาใช้ในการประมาณค่าซึ่งจะก่อให้เกิดความเอนเอียงจากการเลือกตัวอย่าง (sample selection bias) ที่มาจากความผิดพลาดในการระบุแบบจำลอง (specification error) โดยการละทิ้งตัวแปร (omitted variable) และยังนำเสนอ

ถึงวิธีการประมาณค่าที่มีความน่าเชื่อถือในการขจัดความผิดพลาดจากการระบุแบบจำลองในกรณีของตัวอย่างที่ถูกเซนเซอร์

ทั้งนี้ Heckman อธิบายว่า ความเอนเอียงจากการเลือกตัวอย่าง (sample selection bias) อาจเกิดขึ้นจากเหตุผล 2 ประการ คือ ความเอนเอียงที่เกิดจากธรรมชาติของข้อมูลที่ใช้ หรืออาจเกิดจากการเลือกตัวอย่างในการวิเคราะห์ของนักวิจัยเองที่ทำให้ข้อมูลที่ได้มีความเอนเอียง เช่น การหาความสัมพันธ์ระหว่างรายได้ของครัวเรือนกับการใช้จ่ายในสินค้าฟุ่มเฟือย (luxury goods) ดังแสดงในภาพที่ ก.6 ซึ่งข้อมูลที่สังเกตได้จะมีเฉพาะครัวเรือนที่มีการใช้จ่ายมากกว่าศูนย์ ทั้งนี้ การที่บางครัวเรือนไม่เลือกซื้อสินค้าฟุ่มเฟือยอาจเนื่องมาจากการที่รายได้ของครัวเรือนเหล่านั้นอยู่ต่ำกว่าเส้นความยากจน (poverty line) ทำให้ข้อมูลที่สังเกตได้ถูกเซนเซอร์ที่เส้นความยากจน ซึ่ง Heckman กล่าวว่า การประมาณค่าแบบจำลองโดยเลือกใช้เฉพาะข้อมูลค่าใช้จ่ายในสินค้าฟุ่มเฟือยที่สังเกตได้ โดยละเลยข้อมูลที่อยู่ต่ำกว่าเส้นความยากจนซึ่งไม่สามารถสังเกตได้จะทำให้การประมาณค่าที่ได้มีความเอนเอียงที่เกิดจากการละทิ้งตัวแปร

ภาพที่ ก.6

แผนภาพความสัมพันธ์ระหว่างรายได้ของครัวเรือนกับการใช้จ่ายในสินค้าฟุ่มเฟือย



โครงสร้างแบบจำลอง Heckman selection ประกอบด้วยสมการ 2 สมการ คือ สมการทางเลือก (selection equation) และสมการค่าใช้จ่ายหรือสมการถดถอย (regression equation) โดยหลักการของแบบจำลอง Heckman selection เป็นการวางหลักเกณฑ์แบบจำลอง

โทบิทและแบบจำลองตัดปลายใหม่ (generalize tobit and truncated regression model) โดยแสดงถึงกระบวนการในคัดเลือกข้อมูลตัวอย่างที่ถูกเซนเซอร์และไม่ถูกเซนเซอร์ได้อย่างชัดเจน โดยสมมติว่า ค่าใช้จ่ายในการวิจัยและพัฒนา, y*, จะถูกสังเกตได้เฉพาะในบริษัทที่ตัดสินใจทำการวิจัยและพัฒนาเท่านั้น, z* >0, โดยการที่บริษัทจะตัดสินใจทำการวิจัยและพัฒนาหรือไม่นั้นบริษัทจะทำการวิเคราะห์เปรียบเทียบความแตกต่างระหว่างกำไรและต้นทุนที่ได้จากการทำวิจัยและพัฒนา ซึ่งหน่วยผลิตจะตัดสินใจทำการวิจัยและพัฒนาหากกำไรที่ได้มีมากกว่าต้นทุนที่เสียไป แต่เนื่องจากกำไรสุทธิที่แท้จริง , z* , เป็นค่าที่ไม่สามารถสังเกตได้ ทำได้เพียงแต่สังเกตค่า z=1 หากบริษัทตัดสินใจทำการวิจัยและพัฒนา และ z=0 หากบริษัทตัดสินใจไม่ทำการวิจัยและพัฒนาเท่านั้น แสดงโดย

สมการทางเลือก

$$\begin{aligned} z_i^* &= W_i'\gamma + u_i \\ z_i &= 0 \quad if\ z_i^* \le 0 \\ z_i &= 1 \quad if\ z_i^* > 0 \end{aligned} \tag{ก.35}$$

สมการค่าใช้จ่าย

$$\begin{aligned} y_i^* &= X_i'\beta + \varepsilon_i \\ y_i &= y_i^* \quad if\ z_i = 1 \\ y_i &= 0 \quad if\ z_i = 0 \end{aligned} \tag{ก.36}$$

โดยที่

$$\begin{pmatrix} u_i \\ \varepsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix} \right]$$

โดยที่ W_i และ X_i คือ เวกเตอร์ของตัวแปรอธิบายของสมการทางเลือกและสมการค่าใช้จ่าย จะเห็นได้ว่า แบบจำลอง Heckman selection เป็นแบบจำลองที่ประกอบกันขึ้นจากแบบจำลองโพรบิทและโทบิท

ทั้งนี้ จากแบบจำลองโพรบิทจะได้ว่าความน่าจะเป็นที่ z จะมีค่าเท่ากับ 1 คือ

$$Prob(z=1|W_i) = \Phi(W_i'\gamma)$$

และความน่าจะเป็นที่ z=0 คือ

$$Prob(z=0|W_i) = 1-\Phi(W_i'\gamma)$$

โดยที่ $\Phi(W_i'\gamma)$ คือ ฟังก์ชันการแจงแจงความน่าจะเป็นสะสมแบบมาตรฐาน (standard normal cumulative distribution function) และเนื่องจากตัวแปร y_i จะถูกสังเกตได้ก็ต่อเมื่อ z_i^* มีค่ามากกว่าศูนย์ ซึ่งจากทฤษฎี moment of truncated bivariate normal distribution[3] จะได้ว่า

$$\begin{aligned} E(y_i|z_i^*>0) &= E\left[y_i|u_i > -W_i'\gamma\right] \\ &= E\left[X_i'\beta+\varepsilon_i|u_i > -W_i'\gamma\right] \\ &= X_i'\beta + E\left[\varepsilon_i|u_i > -W_i'\gamma\right] \\ E(y_i|z_i^*>0) &= X_i'\beta + \rho\sigma_\varepsilon\lambda_i(\alpha_u) \end{aligned} \qquad \text{(ก.37)}$$

โดยที่ $\alpha_u = -W_i'\gamma/\sigma_u$ และ $\lambda(\alpha_u) = \dfrac{\phi(W_i'\gamma/\sigma_u)}{\Phi(W_i'\gamma/\sigma_u)}$ และเรียก $\lambda(\alpha_u)$ ว่า inverse Mill ratio เราสามารถเขียนสมการถดถอยสำหรับการตัดปลายได้ดังนี้

$$\begin{aligned} y_i|z_i^*>0 &= E\left[y_i|z_i^*>0\right]+v_i \\ &= X_i'\beta+\beta_\lambda\lambda_i(\alpha_u)+v_i \end{aligned} \qquad \text{(ก.38)}$$

โดยที่ $\beta_\lambda = \rho\sigma_\varepsilon$ จากสมการ ก.38 จะเห็นได้ว่า หากประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) ของตัวแปรอธิบาย X และ λ ที่มีต่อ y ค่าประมาณ $\hat{\beta}$ และ $\hat{\beta}_\lambda$ ที่ได้จะมีความน่าเชื่อถือ (consistent) แต่หากละทิ้งตัวแปร λ จะทำให้เกิดปัญหาที่เกิดจากความผิดพลาดจากการระบุแบบจำลอง (specification error) ที่มาจากการละทิ้งตัวแปร ทำให้พารามิเตอร์ที่ประมาณค่าได้จะมีความไม่น่าเชื่อถือ (inconsistent)

[3] อ้างใน Greene (2002) หน้า 781

ทั้งนี้ ในการประมาณค่าแบบจำลอง Heckman selection สามารถทำได้ 2 วิธี คือ การประมาณค่าแบบ maximum likelihood และการประมาณค่าแบบ 2 ขั้นตอน (two-step procedure) ดังนี้

1) การประมาณค่าด้วยวิธี maximum likelihood

จากการที่สมการค่าใช้จ่ายจะถูกสังเกตได้ก็ต่อเมื่อบริษัทตัดสินใจทำการวิจัยและพัฒนา และการที่บริษัทจะตัดสินใจทำการวิจัยและพัฒนาก็ต่อเมื่อกำไรสุทธิที่แท้จริงมากกว่าศูนย์ จึงทำให้แบบจำลอง Heckman selection ประกอบด้วยสมการ likelihood ที่มาจากแบบจำลองโพรบิทและแบบจำลองแบบตัดปลาย โดยที่สมการ likelihood ของแบบจำลองโพรบิทแสดงโดย

$$L_{probit} = \prod_{z_i=0}\left(1-\Phi\left(W_i'\gamma\right)\right)\prod_{z_i=1}\Phi\left[\frac{W_i'\gamma+\rho(y_i-X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$\ln L_{probit} = \sum_{z_i=0}\ln\left(1-\Phi\left(W_i'\gamma\right)\right)+\sum_{z_i=1}\ln\Phi\left[\frac{W_i'\gamma+\rho(y_i-X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]$$

และสมการ likelihood สำหรับกลุ่มตัวอย่างที่ไม่ถูกเซนเซอร์สามารถเขียนได้ ดังนี้

$$L_{truncate} = \prod_{z_i=1}\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}e^{-\frac{(y_i-X_i'\beta)^2}{2\sigma_\varepsilon^2}}$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$\ln L_{truncate} = \sum_{z_i=1}-\frac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right]-\frac{\left(y_i-X_i'\beta\right)^2}{2\sigma_\varepsilon^2}$$

รวมสมการทั้งสองเข้าด้วยกันจะได้ likelihood function สำหรับแบบจำลอง Heckman selection คือ

$$L=\begin{cases}\prod_{z_i=1}\Phi\left[\dfrac{W_i'\gamma+\rho(y_i-X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]\cdot\prod_{z_i=1}\dfrac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}e^{-\frac{(y_i-X_i'\beta)^2}{2\sigma_\varepsilon^2}} & if\ \ y_i\ observed\\ \prod_{z_i=0}\left(1-\Phi\left(W_i'\gamma\right)\right) & if\ \ y_i\ not\ observed\end{cases}$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$L=\begin{cases}\sum_{z_i=1}\ln\Phi\left[\dfrac{W_i'\gamma+\rho(y_i-X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]+\sum_{z_i=1}-\dfrac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right]-\dfrac{\left(y_i-X_i'\beta\right)^2}{2\sigma_\varepsilon^2} & if\ \ y_i\ observed\\ \sum_{z_i=0}\ln\left(1-\Phi\left(W_i'\gamma\right)\right) & if\ \ y_i\ not\ observed\end{cases}$$

แก้สมการ log likelihood ภายใต้เงื่อนไขลำดับที่หนึ่ง (first order condition) เพื่อหาค่าพารามิเตอร์ไม่ทราบค่า $\hat{\gamma},\hat{\beta},\hat{\rho},\hat{\sigma}$ ซึ่งเป็นตัวประมาณของพารามิเตอร์ γ,β,ρ,σ ตามลำดับ อย่างไรก็ตาม ในการประมาณค่าแบบ maximum likelihood ไม่สามารถประมาณค่าพารามิเตอร์ ρ,σ โดยตรงได้ ทำได้เพียงแต่ประมาณค่า $\ln\sigma$ และค่า inverse hyperbolic tangent of ρ โดยที่

$$atanh\,\rho=\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

ทั้งนี้ ในการตรวจสอบผลกระทบที่มาจากการเลือกตัวอย่าง (selectivity effect) นักเศรษฐศาสตร์ส่วนใหญ่จะดูที่ค่า $\lambda=\rho\sigma_\varepsilon$ โดยเรียกว่า nonselection hazard

2. การประมาณค่าแบบ 2 ขั้นตอน

ในบางครั้งที่ขนาดของข้อมูลมีจำนวนมาก การประมาณค่าด้วยวิธี maximum likelihood จำเป็นต้องใช้เวลามาก ซึ่งการประมาณค่าแบบ 2 ขั้นตอน (บางครั้งเรียกว่า Heckman

two-step procedure) จะมีความสะดวกกว่าในการจัดการกับข้อมูลที่มีขนาดใหญ่ โดยขั้นตอนการประมาณค่าด้วยวิธี Heckman two step มีดังนี้

ขั้นตอนแรก จะประมาณค่าสมการทางเลือกด้วยวิธี maximum likelihood เหมือนในแบบจำลองโพรบิท เพื่อประมาณค่าพารามิเตอร์ $\hat{\gamma}$ ซึ่งแต่ละค่าของตัวอย่างจะสามารถคำนวณค่า inverse Mill ratio $\hat{\lambda}(\alpha_u) = \frac{\phi(W_i'\hat{\gamma})}{\Phi(W_i'\hat{\gamma})}$ และค่า $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i - W_i\hat{\gamma})$ ออกมาได้

ขั้นตอนที่สอง เป็นการประมาณค่าสมการค่าใช้จ่ายหรือสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) ของตัวแปรอธิบาย X และ $\hat{\lambda}$ (ที่คำนวณได้จากสมการทางเลือก) ที่มีต่อ y เพื่อหาค่าประมาณ $\hat{\beta}$ และ $\hat{\beta}_\lambda$ ซึ่งการกำหนดให้ค่า inverse Mill ratio เป็นตัวแปรอธิบายที่รวมอยู่ในสมการค่าใช้จ่ายจะทำให้สามารถขจัดปัญหาที่เกิดจากการเลือกตัวอย่างได้ (selectivity bias)

สำหรับการแปลความหมายของแบบจำลอง Heckman selection จะอธิบายโดยใช้ค่าผลกระทบส่วนเพิ่ม (marginal effect) เช่นเดียวกับในแบบจำลองโพรบิทและโทบิท และเนื่องจากขั้นตอนที่สองของแบบจำลอง Heckman two step เป็นการประมาณค่าด้วยวิธีกำลังสองน้อยที่สุด (OLS) ทำให้ค่าประมาณของค่าสัมประสิทธิ์ที่ได้สามารถแปลผลได้เหมือนกับแบบจำลองสมการถดถอยเชิงเส้นทั่วไป

อย่างไรก็ตาม ถึงแม้ว่าแบบจำลอง Heckman selection จะมีความคล้ายคลึงกับแบบจำลองโทบิทในเรื่องของการเซนเซอร์ข้อมูล แต่จะมีความแตกต่างในเรื่องของการคำนวณค่าความน่าจะเป็นที่ค่าที่สังเกตได้จะถูกเซนเซอร์ (probability for limit observation) และค่าประมาณของตัวแปรตาม โดยในแบบจำลองโทบิทจะคำนวณค่าความน่าจะเป็นที่ค่าที่สังเกตได้จะถูกเซนเซอร์จากเวกเตอร์ของตัวแปรอธิบายและค่าสัมประสิทธิ์ชุดเดียวกัน, $X'\beta$, ขณะที่แบบจำลอง Heckman selection จะคำนวณจากเวกเตอร์ของตัวแปรอธิบายและค่าสัมประสิทธิ์ของสมการทางเลือก $W'\gamma$ ทำให้ความน่าจะที่ข้อมูลจะถูกเซนเซอร์ที่ได้มีความแตกต่างกัน และเนื่องจากค่าประมาณของตัวแปรตามเป็นฟังก์ชันของความน่าจะที่ข้อมูลจะถูกเซนเซอร์และค่าเฉลี่ยของการตัดปลาย ซึ่งหากความน่าจะที่ข้อมูลจะถูกเซนเซอร์มีความแตกต่างกัน ค่าของตัวแปรตามที่ได้ก็จะมีความแตกต่างกันด้วย และอีกประการหนึ่งคือ ในแบบจำลอง sample selection ค่าที่สังเกตได้จะถูกเซนเซอร์จากตัวแปรตามของอีกสมการหนึ่ง ทำให้สามารถเข้าใจถึงกระบวนการเซนเซอร์ได้อย่างชัดเจนและมีความสมเหตุสมผล ในขณะที่แบบจำลองโทบิทไม่สามารถอธิบายเหตุผลว่าทำไมข้อมูลถึงถูกเซนเซอร์

5. <u>แบบจำลองสมการเกี่ยวพันธ์กัน (simultaneous equation model)</u>

ระบบสมการแบบเกี่ยวพันธ์กันเกิดขึ้นเมื่อตัวแปรตาม y ถูกกำหนดขึ้นจากตัวแปรอธิบาย X ในขณะเดียวกันตัวแปรอธิบายบางตัวของ X ก็ถูกกำหนดจากตัวแปรตาม y หรืออาจกล่าวได้ว่าตัวแปรตาม y มีความสัมพันธ์แบบเกี่ยวพันธ์กันกับตัวแปร X ซึ่งการประมาณค่าแบบจำลองลักษณะนี้จำเป็นต้องทำการพิจารณาพร้อมกัน เพื่อความเข้าใจในแบบจำลองสมการแบบเกี่ยวพันธ์กัน พิจารณาสมการทั่วไปของแบบจำลอง แสดงโดย

$$y_1 = \gamma_1 y_2 + \beta_1' X_1 + \varepsilon_1 \qquad \text{(ก.39)}$$

$$y_2 = \gamma_2 y_1 + \beta_2' X_2 + \varepsilon_2 \qquad \text{(ก.40)}$$

จะเห็นได้ว่า ตัวแปรตาม y_1 และ y_2 มีความสัมพันธ์ต่อกันและกัน (mutually dependent) โดยตัวแปรภายใน y_1 เป็นตัวแปรอธิบายในสมการ ก.40 ในขณะเดียวกันตัวแปรภายใน y_1 ก็เป็นตัวแปรอธิบายในสมการ ก.39 ซึ่งในสถานการณ์เช่นนี้ทำให้ตัวแปรตามมีความสัมพันธ์กับค่าความคลาดเคลื่อน (error term) โดยตัวแปรตาม y_1 มีความสัมพันธ์กับ ε_2, $\text{cov}(y_1, \varepsilon_2) \neq 0$ เช่นเดียวกับตัวแปรตาม y_2 ที่มีความสัมพันธ์กับ ε_1, $\text{cov}(y_2, \varepsilon_1) \neq 0$ นอกจากนี้ การที่ค่าความคลาดเคลื่อน $\varepsilon_1, \varepsilon_2$ มีการกระจายตัวของข้อมูล (nonstochastic) ทำให้ตัวแปรภายใน y_1 และ y_2 มีการกระจายตัวของข้อมูลด้วย ซึ่งการที่ตัวแปรภายในที่เป็นตัวแปรอธิบายในอีกสมการหนึ่งมีความสัมพันธ์กับค่าความคลาดเคลื่อนทำให้ผลการประมาณค่าที่ได้จากวิธีกำลังสองน้อยที่สุด (OLS) มีความเอนเอียงและไม่น่าเชื่อถือ (bias and inconsistent) ทั้งนี้เนื่องมาจาก สมมติฐานของการประมาณค่าด้วยวิธีกำลังสองน้อยที่สุดตัวแปรอธิบายจำเป็นต้องไม่มีการกระจายตัวของข้อมูลและไม่มีความสัมพันธ์ (เป็นอิสระ) กับค่าความคลาดเคลื่อน เราเรียกปัญหาในลักษณะนี้ว่า “ความเอนเอียงที่มาจากสมการแบบเกี่ยวพันธ์กัน (simultaneity bias)”

ในการแก้ปัญหาความเอนเอียงที่เกิดจากสมการแบบเกี่ยวพันธ์กันโดยทั่วไปจะนิยมใช้วิธีตัวแปรเครื่องมือ (instrument variable) หรือวิธี Indirect least square และวิธีการประมาณค่าแบบสองขั้นตอนน้อยที่สุด (two-stage least square) แต่เนื่องด้วยข้อจำกัดของข้อมูลที่สังเกตได้โดยข้อมูลผลิตภาพการผลิตมีลักษณะข้อมูลแบบต่อเนื่อง (continuous data) ขณะที่ข้อมูลผลผลิตทางนวัตกรรม (innovation output) มีลักษณะข้อมูลแบบสองทางเลือก (dichotomous

data) ทำให้ระบบสมการแบบเกี่ยวพันธ์กันที่ใช้ในการศึกษามีลักษณะแบบตัวแปรภายในตัวหนึ่งเป็นข้อมูลแบบต่อเนื่องและอีกตัวหนึ่งมีลักษณะข้อมูลแบบสองทางเลือก ซึ่งการที่ตัวแปรภายในมีลักษณะแบบสองทางเลือก ทำให้การประมาณค่าโดยวิธีกำลังสองน้อยที่สุด (least square) มีความเอนเอียงและไม่น่าเชื่อถือ ทั้งนี้ Maddala (1983) ได้เสนอแบบจำลอง Treatment effects model ที่มีความเหมาะสมในการประมาณค่าแบบจำลองสมการแบบเกี่ยวพันธ์กันในกรณีที่ตัวแปรภายในตัวหนึ่งมีลักษณะข้อมูลแบบสองทางเลือก โดยพิจารณาถึงผลกระทบของตัวแปรภายในที่มีลักษณะแบบสองทางเลือก (endogenously chosen binary) ที่ส่งผลต่อตัวแปรภายในที่มีความต่อเนื่องของข้อมูล (endogenous continuous variable)

โครงสร้างแบบจำลอง treatment effects ประกอบด้วยสมการ 2 สมการ คือ สมการถดถอย (regression equation) และสมการส่งผลกระทบ (equation for treatment effects) โดยสมมติว่า ฟังก์ชันการผลิตความรู้ (knowledge production function) (ผลผลิตทางนวัตกรรม) ซึ่งเป็นผลจากปัจจัยนำเข้าทางด้านนวัตกรรม (innovation input) ส่งผลต่อผลิตภาพการผลิตของหน่วยผลิต โดยที่ผลผลิตทางนวัตกรรมมีลักษณะข้อมูลเป็นตัวแปรหุ่น (dummy variable) โดยเท่ากับ 1 เมื่อหน่วยผลิตมีนวัตกรรมผลิตภัณฑ์และ/หรือนวัตกรรมกระบวนการอยู่แล้วในช่วงเวลาที่ทำการสำรวจ และเท่ากับ 0 หากหน่วยผลิตไม่มีนวัตกรรมผลิตภัณฑ์และ/หรือนวัตกรรมกระบวนการ ขณะที่ผลิตภาพการผลิตมีลักษณะข้อมูลแบบต่อเนื่อง โดยวัดจากลอการิทึมของยอดขายทั้งหมดต่อจำนวนแรงงาน โดยมีรูปแบบสมการ ดังนี้

สมการถดถอย

$$y_i = X_i'\beta + \delta z_i + \varepsilon_i \qquad \text{(ก.41)}$$

สมการ treatment effects

$$\begin{aligned} z_i^* &= W_i'\gamma + u_i \\ z_i &= \begin{cases} 1 & if\ z_i^* > 0 \\ 0 & if\ z_i^* \le 0 \end{cases} \end{aligned} \qquad \text{(ก.42)}$$

โดยที่

$$\begin{pmatrix} u_i \\ \varepsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix} \right]$$

โดยที่ W_i และ X_i คือ เวกเตอร์ของตัวแปรอธิบายของสมการ treatment effects และสมการถดถอย ตามลำดับ และความคลาดเคลื่อน ε_i, u_i มีการแจกแจงแบบปกติและมีค่า

สหสัมพันธ์ ρ (bivariate normal and correlation ρ) ดังนั้น ค่าคาดหวังแบบมีเงื่อนไข (conditional expected) เมื่อหน่วยผลิตมีนวัตกรรมผลิตภัณฑ์และ/หรือนวัตกรรมกระบวนการ คือ

$$\begin{aligned} E(y_i|z_i=1,X_i,W_i) &= X_i'\beta+\delta+E\left[\varepsilon_i|z_i=1,X_i,W_i\right] \\ &= X_i'\beta+\delta+\rho\sigma_\varepsilon\left(\frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)}\right) \\ &= X_i'\beta+\delta+\rho\sigma_\varepsilon\lambda\left(-W_i'\gamma\right) \end{aligned} \qquad \text{(ก.43)}$$

ขณะที่ ค่าคาดหวังแบบมีเงื่อนไขเมื่อหน่วยผลิตไม่มีนวัตกรรมผลิตภัณฑ์และ/หรือนวัตกรรมกระบวนการ คือ

$$\begin{aligned} E(y_i|z_i=0,X_i,W_i) &= X_i'\beta+E\left[\varepsilon_i|z_i=0,X_i,W_i\right] \\ &= X_i'\beta+\rho\sigma_\varepsilon\left(\frac{-\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)}\right) \end{aligned} \qquad \text{(ก.44)}$$

โดยที่ $\Phi(W_i'\gamma)$ คือ ฟังก์ชันการแจงแจงความน่าจะเป็นสะสมแบบมาตรฐาน (standard normal cumulative distribution function) ทั้งนี้ ความแตกต่างของค่าคาดหวังแบบมีเงื่อนไขทั้งสองแบบคือ

$$\begin{aligned} E(y_i|z_i=1,X_i,W_i)-E(y_i|z_i=0,X_i,W_i) &= \delta+\rho\sigma_\varepsilon\left(\frac{\phi(\cdot)}{\Phi(\cdot)}+\frac{\phi(\cdot)}{1-\Phi(\cdot)}\right) \\ &= \delta+\rho\sigma_\varepsilon\left(\frac{\phi(\cdot)\left(1-\Phi(\cdot)\right)+\phi(\cdot)\Phi(\cdot)}{\Phi(\cdot)\left(1-\Phi(\cdot)\right)}\right) \\ E(y_i|z_i=1,X_i,W_i)-E(y_i|z_i=0,X_i,W_i) &= \delta+\rho\sigma_\varepsilon\left(\frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)\left(1-\Phi\left(W_i'\gamma\right)\right)}\right) \end{aligned} \qquad \text{(ก.45)}$$

เราสามารถเขียนสมการถดถอยแบบมีเงื่อนไข ได้ดังนี้

$$\begin{aligned} y_i \mid z_i &= X_i'\beta + \delta z_i + \rho\sigma_\varepsilon \lambda_i + v_i \\ &= X_i'\beta + \delta z_i + \beta_\lambda \lambda_i + v_i \end{aligned} \qquad \text{(ก.46)}$$

โดยที่ $\beta_\lambda = \rho\sigma_\varepsilon$ จากสมการ ก.46 จะเห็นได้ว่า หากประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) โดยละทิ้งตัวแปร λ เท่ากับว่าเราเพิกเฉยต่อการประมาณค่าสัมประสิทธิ์ที่มาจากตัวแปรส่งผลกระทบ (treatment dummy variable) ในสมการ ก.45 จะทำให้ผลการประมาณค่าที่ได้เกิดปัญหาจากความผิดพลาดในการระบุแบบจำลอง (specification error) ที่มาจากการละทิ้งตัวแปร ทำให้พารามิเตอร์ที่ประมาณค่าได้มีความเอนเอียงและไม่น่าเชื่อถือ

สำหรับการประมาณค่าแบบจำลอง treatment effects Maddala (1983) ได้นำเสนอวิธีการประมาณค่าทั้งวิธี maximum likelihood และวิธีการประมาณค่า 2 ขั้นตอน (two-step estimator) ดังนี้

1. การประมาณค่าด้วยวิธี maximum likelihood

จากการที่แบบจำลอง treatment effects เป็นแบบจำลองที่พิจารณาถึงผลกระทบของตัวแปรภายในที่มีลักษณะแบบสองทางเลือกที่ส่งผลต่อตัวแปรภายในที่มีความต่อเนื่องของข้อมูล จึงทำให้สมการ likelihood function ของแบบจำลองประกอบด้วยสมการ likelihood ที่มาจากแบบจำลองสมการถดถอยทั่วไปและแบบจำลองสองทางเลือก (โพรบิท) โดยที่สมการ likelihood ของแบบจำลองโพรบิทแสดงโดย

$$L_{probit} = \prod_{z_i=0}\left(1-\Phi\left[\frac{W_i'\gamma + \rho(y_i - X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]\right)\bullet\prod_{z_i=1}\Phi\left[\frac{W_i'\gamma + \rho(y_i - X_i'\beta - \delta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$\ln L_{probit} = \sum_{z_i=0}\ln\left(1-\Phi\left[\frac{W_i'\gamma + \rho(y_i - X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]\right) + \sum_{z_i=1}\ln\Phi\left[\frac{W_i'\gamma + \rho(y_i - X_i'\beta - \delta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right]$$

และสมการ likelihood สำหรับสมการถดถอยทั่วไปสามารถเขียนได้ ดังนี้

$$L_{normal} = \prod_{z_i=0} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{-\frac{(y_i - X_i'\beta)^2}{2\sigma_\varepsilon^2}} \bullet \prod_{z_i=1} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{-\frac{(y_i - X_i'\beta - \delta)^2}{2\sigma_\varepsilon^2}}$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$\ln L_{normal} = \sum_{z_i=0} -\frac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right] - \frac{\left(y_i - X_i'\beta\right)^2}{2\sigma_\varepsilon^2} + \sum_{z_i=1} -\frac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right] - \frac{\left(y_i - X_i'\beta - \delta\right)^2}{2\sigma_\varepsilon^2}$$

รวมสมการทั้งสองเข้าด้วยกันจะได้ likelihood function สำหรับแบบจำลอง Treatment effects คือ

$$L = \begin{cases} \prod_{z_i=1} \Phi\left[\dfrac{W_i'\gamma + \rho(y_i - X_i'\beta - \delta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right] \bullet \prod_{z_i=1} \dfrac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{-\frac{(y_i - X_i'\beta - \delta)^2}{2\sigma_\varepsilon^2}} & if \ z_i = 1 \\ \prod_{z_i=0} \Phi\left[\dfrac{W_i'\gamma + \rho(y_i - X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right] \bullet \prod_{z_i=0} \dfrac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{-\frac{(y_i - X_i'\beta)^2}{2\sigma_\varepsilon^2}} & if \ z_i = 0 \end{cases}$$

เมื่อแปลงให้อยู่ในรูปของลอการิทึมธรรมชาติ (natural logarithm) จะได้ว่า log likelihood คือ

$$L = \begin{cases} \sum_{z_i=1} \ln\Phi\left[\dfrac{W_i'\gamma + \rho(y_i - X_i'\beta - \delta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right] + \sum_{z_i=1} -\dfrac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right] - \dfrac{\left(y_i - X_i'\beta - \delta\right)^2}{2\sigma_\varepsilon^2} & if \ z_i = 1 \\ \sum_{z_i=0} \ln\Phi\left[\dfrac{W_i'\gamma + \rho(y_i - X_i'\beta)/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right] + \sum_{z_i=0} -\dfrac{1}{2}\left[\ln\left(2\pi\sigma_\varepsilon^2\right)\right] - \dfrac{\left(y_i - X_i'\beta\right)^2}{2\sigma_\varepsilon^2} & if \ z_i = 0 \end{cases}$$

แก้สมการ log likelihood ภายใต้เงื่อนไขลำดับที่หนึ่ง (first order condition) เพื่อหาค่าพารามิเตอร์ไม่ทราบค่า $\hat{\gamma}, \hat{\beta}, \hat{\rho}, \hat{\sigma}, \hat{\delta}$ ซึ่งเป็นตัวประมาณของพารามิเตอร์ $\gamma, \beta, \rho, \sigma, \delta$

ตามลำดับ อย่างไรก็ตาม ในการประมาณค่าแบบ maximum likelihood ไม่สามารถประมาณค่าพารามิเตอร์ ρ, σ โดยตรงได้ ทำได้เพียงแต่ประมาณค่า $\ln \sigma$ และค่า inverse hyperbolic tangent of ρ โดยที่

$$atanh\, \rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$$

ทั้งนี้ ในการตรวจสอบผลกระทบที่มาจากการเลือกตัวอย่าง (selectivity effect) นักเศรษฐศาสตร์ส่วนใหญ่จะดูที่ค่า $\lambda = \rho \sigma_{\varepsilon}$ โดยเรียกว่า nonselection hazard

2. การประมาณค่าแบบ 2 ขั้นตอน

ในบางครั้งที่ขนาดของข้อมูลมีจำนวนมาก การประมาณค่าด้วยวิธี maximum likelihood จำเป็นต้องใช้เวลามาก ซึ่งการประมาณค่าแบบ 2 ขั้นตอน (two-step estimator) จะมีความสะดวกกว่าในการจัดการกับข้อมูลที่มีขนาดใหญ่ โดยขั้นตอนการประมาณค่าด้วยวิธี two step estimator มีดังนี้

ขั้นตอนแรก จะประมาณค่าสมการ treatment effects ด้วยวิธี maximum likelihood เหมือนในแบบจำลองโพรบิท เพื่อประมาณค่าพารามิเตอร์ไม่ทราบค่า $\hat{\gamma}$ ซึ่งเป็นตัวประมาณค่าของพารามิเตอร์ γ เมื่อแทนค่าพารามิเตอร์ไม่ทราบค่าที่ได้จากการประมาณจะได้สมการประมาณค่าแบบจำลองโพรบิท ดังนี้

$$Prob\left(z_i = 1 \middle| W_i \right) = \Phi\left(W_i' \hat{\gamma} \right)$$

แต่ละค่าของตัวอย่างจะสามารถคำนวณค่า inverse Mill ratio ได้ โดยที่

$$\hat{\lambda}_i = \begin{cases} \dfrac{\phi(W_i'\hat{\gamma})}{\Phi(W_i'\hat{\gamma})} & if\ z_i = 1 \\ \dfrac{-\phi(W_i'\hat{\gamma})}{1-\Phi(W_i'\hat{\gamma})} & if\ z_i = 0 \end{cases}$$

และค่า $\hat{d}_i = \hat{\lambda}_i (\hat{\lambda}_i - W_i'\hat{\gamma})$

ขั้นตอนที่สอง เป็นการประมาณค่าสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) ของตัวแปรอธิบาย X δ และ $\hat{\lambda}$ (ที่คำนวณได้จากสมการ treatment effects) ที่มีต่อ y เพื่อหาค่าประมาณ $\hat{\beta}$ และ $\hat{\beta}_{\lambda}$ โดยมีสมการถดถอยที่ใช้ในการประมาณค่าคือ

$$\begin{aligned} y_i \left| z_i \right. &= E(y_i \left| z_i \right.) + \varepsilon_i \\ &= X_i'\beta + \delta z_i + \beta_\lambda \hat{\lambda}_i + \varepsilon_i \end{aligned}$$

และ $$\operatorname{var}\left(y_i \left| z_i \right. \right) = \sigma_\varepsilon^2 \left(1 - \rho^2 d_i \right)$$

ทั้งนี้ การกำหนดให้ค่า inverse Mill ratio เป็นตัวแปรอธิบายที่รวมอยู่ในสมการถดถอยแล้วทำการประมาณค่าด้วยวิธีกำลังสองน้อยที่สุด (OLS) ของตัวแปรอธิบาย X δ และ $\hat{\lambda}$ ที่มีต่อ y จะทำให้สามารถขจัดปัญหาความผิดพลาดจากการระบุแบบจำลองได้ สำหรับการแปลความหมายของแบบจำลอง treatment effects จะอธิบายโดยใช้ค่าผลกระทบส่วนเพิ่ม (marginal effect) เช่นเดียวกับในแบบจำลอง Heckman selection