บทที่ 2
ทบทวนการวิเคราะห์การถดถอยเชิงเส้น

2.1 กล่าวนำ

การวิเคราะห์การถดถอย (regression analysis) เป็นวิธีการทางสถิติที่ศึกษาถึงความสัมพันธ์เชิงสถิติ (statistical relation) ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยการศึกษาถึงความสัมพันธ์เชิงสถิติดังกล่าว เป็นไปในลักษณะที่ต้องการจะประมาณค่าของตัวแปรหนึ่ง ซึ่งเรียกว่าตัวแปรตาม (dependent variable) จากตัวแปรอื่นๆ ที่เหลือ ซึ่งเรียกว่าตัวแปรอิสระ (independent variables) และการแสดงความสัมพันธ์เชิงสถิติระหว่างตัวแปรตามกับตัวแปรอิสระ จะแสดงในรูปโมเดลทางคณิตศาสตร์ (mathematical models) ค่าประมาณของตัวแปรตามจากโมเดล จะอยู่ในรูปค่าเฉลี่ย (mean) หรือค่าคาดหวัง (expected value) เช่น

ถ้า X หมายถึงค่าใช้จ่ายในการโฆษณาสินค้า และ Y หมายถึงมูลค่าสินค้าที่ขายได้ และรู้ว่า X มีความสัมพันธ์เชิงสถิติกับ Y โมเดลแสดงความสัมพันธ์ของข้อมูลดังกล่าวเป็น $\hat{Y} = 9.5 + 2.1X$

ในลักษณะเช่นนี้ X ก็คือตัวแปรอิสระที่มีอิทธิพลต่อตัวแปรตาม Y สามารถประมาณค่า Y ได้จากโมเดลนี้ โดยค่าประมาณที่ได้จะคือค่าเฉลี่ยของ Y ณ ค่าของ X ที่กำหนด เป็นต้น

โมเดลทางคณิตศาสตร์ ที่แสดงความสัมพันธ์เชิงสถิติระหว่างตัวแปรตามกับตัวแปรอิสระต่างๆ อาจเป็นแบบเชิงเส้น (linear) หรือแบบไม่ใช่เชิงเส้น (non linear) ก็ได้ ในที่นี้จะกล่าวเฉพาะความสัมพันธ์แบบเชิงเส้นเท่านั้น

2.2 โมเดลการถดถอยเชิงเส้น (Linear regression model)

โมเดลการถดถอยเชิงเส้น เป็นโมเดลทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่าง

ตัวแปรตามกับตัวแปรอิสระอยู่ในรูปของสมการเชิงเส้น ซึ่งเรียกว่าสมการถดถอยเชิงเส้น แบ่งออกเป็น 2 ประเภทคือ

1. สมการถดถอยเชิงเส้นอย่างง่าย (simple linear regression equation)
2. สมการถดถอยเชิงเส้นพหุ (multiple linear regression equation)

2.2.1 โมเดลการถดถอยเชิงเส้นอย่างง่าย

โมเดลทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัว คือ ตัวแปรตามและตัวแปรอิสระ ในรูปของสมการเชิงเส้น จะได้สมการถดถอยเป็นดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad ; \quad i = 1, 2, \ldots, N \qquad (2.2.1)$$

$$Y_i = b_0 + b_1 X_i + e_i \quad ; \quad i = 1, 2, \ldots, n \qquad (2.2.2)$$

สมการ (2.2.1) เป็นสมการสำหรับประชากร (population) และสมการ (2.2.2) เป็นสมการสำหรับตัวอย่าง (sample)

เมื่อ Y_i คือตัวแปรตาม X_i คือตัวแปรอิสระ β_0, b_0 และ β_1, b_1 คือค่าคงที่ และสัมประสิทธิ์การถดถอย (regression coefficient) ของประชากรและของตัวอย่าง ตามลำดับ ϵ_i และ e_i คือความคลาดเคลื่อนเชิงสุ่ม(random error) ของประชากรและของตัวอย่าง ตามลำดับ

2.2.2 โมเดลการถดถอยเชิงเส้นพหุ

โมเดลทางคณิตศาสตร์ ที่แสดงความสัมพันธ์ระหว่างตัวแปรตาม 1 ตัว กับตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป ในรูปของสมการเชิงเส้น จะได้สมการถดถอยเป็นดังนี้

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{q-1} X_{q-1\,i} + \epsilon_i \qquad (2.2.3)$$

$$= \sum_{j=0}^{q-1} \beta_j X_{ji} + \epsilon_i \quad ; \quad i = 1, 2, \ldots, N \; ; \; X_{0i} = 1$$

$$Y_i = b_o + b_1X_{1i} + b_2X_{2i} + \dots + b_{q-1}X_{q-1i} + e_i \qquad (2.2.4)$$

$$= \sum_{j=0}^{q-1} b_jX_{ji} + e_i \qquad ; \quad i = 1, 2, \dots, n$$

สมการ (2.2.3) เป็นสมการสำหรับประชากร สมการ (2.2.4) เป็นสมการสำหรับตัวอย่าง

เมื่อ Y_i คือตัวแปรตาม X_{ji} คือตัวแปรอิสระ β_o, b_o และ β_j, b_j คือค่าคงที่ และสัมประสิทธิ์การถดถอย ของประชากรและของตัวอย่าง ตามลำดับ ϵ_i และ e_i คือความคลาดเคลื่อนเชิงสุ่มของประชากรและของตัวอย่าง ตามลำดับ

2.2.3 โมเดลการถดถอยเชิงเส้นในรูปแมทริกซ์

การวิเคราะห์การถดถอย คือการสร้างสมการการถดถอยขึ้นมา เพื่อใช้สำหรับประมาณค่าของตัวแปรตามจากสมการ การจะสร้างสมการดังกล่าวขึ้นมาได้ จำเป็นต้องอาศัยข้อมูลระหว่าง Y_i และ X_{ji} ต่าง ๆ จากตัวอย่างจำนวน n ชุด เพื่อใช้ในการสร้างค่า b_j ซึ่งจะเป็นค่าประมาณของพารามิเตอร์ β_j เพื่อให้เกิดความสะดวกต่อการศึกษาถึงวิธีการต่าง ๆ ของการประมาณค่าพารามิเตอร์ β_j ตลอดจนรายละเอียดต่าง ๆ เกี่ยวกับการวิเคราะห์การถดถอย จะเขียนสมการถดถอย (2.2.4) ให้อยู่ในรูปแมทริกซ์ ดังนี้

ให้ $\underset{nx1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$, $\underset{nxq}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1q-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2q-1} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & X_{n1} & X_{n2} & \dots & X_{nq-1} \end{bmatrix}$

$$\underset{qx1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_{q-1} \end{bmatrix} \quad , \quad \underset{nx1}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

จะได้ $$Y = X\beta + \epsilon \qquad (2.2.5)$$

เป็นสมการถดถอยเชิงเส้นพหุในรูปแมทริกซ์ สำหรับประชากร

และได้ $$Y = Xb + e \qquad (2.2.6)$$

เป็นสมการถดถอยเชิงเส้นพหุในรูปแมทริกซ์สำหรับตัวอย่าง เมื่อมี

$$\underset{qx1}{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_{q-1} \end{bmatrix} \quad \text{และ} \quad \underset{nx1}{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

2.2.4 ข้อสมมติเบื้องต้นในการวิเคราะห์การถดถอย

ในการประมาณค่าพารามิเตอร์ β_j ที่จะได้กล่าวในหัวข้อต่อไป มีข้อสมมติเบื้องต้น (assumptions) ที่สำคัญ ซึ่งควรกล่าวถึงดังนี้ คือ

1. ความคลาดเคลื่อนเชิงสุ่ม ϵ_i มีค่าเฉลี่ย = 0 ; $E(\epsilon_i) = 0$

2. ความแปรปรวน (variance) ของ ϵ_i มีค่าคงที่ ; $V(\epsilon_i) = \sigma^2$

3. ความแปรปรวนร่วม (covariance) ระหว่าง ϵ_i กับ ϵ_j เมื่อ $i \neq j$ มีค่าเท่ากับ 0 ; $Cov(\epsilon_i,\epsilon_j) = 0$ ซึ่งหมายความว่า ϵ_i กับ ϵ_j เมื่อ $i \neq j$ ไม่มีความสัมพันธ์กัน (nocorrelation)

4. การแจกแจงของ ϵ_i มีการแจกแจงแบบปกติ (normal distribution)

5. ตัวแปรอิสระ X_{ji} มีลักษณะเป็นค่าคงที่ คือถูกวัดโดยไม่มีความคลาดเคลื่อน

และจากสมการ (2.2.3) $$Y_i = \sum_{j=0}^{q-1} \beta_j X_{ji} + \epsilon_i$$

จะได้ $$E(Y_i) = \sum_{j=0}^{q-1} \beta_j X_{ji} \qquad (2.2.7)$$

ซึ่งหมายความว่า ณ ค่าของ X_{ji} ที่กำหนด จะได้ค่าเฉลี่ยของ Y_i มีค่าเท่ากับ $\sum_{j=0}^{q-1} \beta_j X_{ji}$

2.3 การประมาณค่าพารามิเตอร์ (Estimation of parameter)

การประมาณค่าพารามิเตอร์ β_j เพื่อใช้สร้างสมการถดถอย จะประมาณได้จากข้อมูลของตัวอย่าง ซึ่งจำเป็นต้องทราบก่อนว่าตัวประมาณ (estimators) ของ β_j คืออะไร

โดยปกติ ตัวประมาณที่ดี ควรมีคุณสมบัติ ดังนี้ คือ

1. เป็นตัวประมาณที่ไม่เอนเอียง (unbiased estimator)

2. เป็นตัวประมาณที่มีความแปรปรวนต่ำสุด (minimum variance estimator)

3. เป็นตัวประมาณที่มีความคงเส้นคงวา (consistency estimator)

4. เป็นตัวประมาณที่มีความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (minimum mean-squared-error estimator)

5. เป็นตัวประมาณที่มีประสิทธิภาพ (efficiency estimator)

6. เป็นตัวประมาณที่มีความพอเพียง (sufficiency estimator)

ในการวิเคราะห์การถดถอย วิธีการที่นิยมใช้ในการหาตัวประมาณที่เหมาะสมของพารามิเตอร์ β_j ได้แก่

1. วิธีกำลังสองต่ำสุด (The least square method)

2. วิธีภาวะน่าจะเป็นสูงสุด (The maximum likelioad method)

2.3.1 การประมาณค่าโดยวิธีกำลังสองต่ำสุด

หลักการของการหาตัวประมาณโดยวิธีกำลังสองต่ำสุด ก็คือ การทำให้ผลรวมของความคลาดเคลื่อนกำลังสองมีค่าต่ำสุด ซึ่งมีขั้นตอนดังนี้ คือ

1. หาค่าความคลาดเคลื่อน ϵ_i

จากสมการ (2.2.5) $Y = X\beta + \epsilon$

$$\epsilon = Y - X\beta$$

2. หาผลรวมของความคลาดเคลื่อนกำลังสอง (sum of squared of error)

$$\epsilon'\epsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta X'Y + \beta'X'X\beta$$

3. หาอนุพันธ์ (differentiated) ของ $\epsilon'\epsilon$ เทียบกับ β แล้วทำให้สมการที่ได้มีค่าเท่ากับ 0 และเมื่อให้ $\hat{\beta} = b$ เป็นค่าประมาณของ β

$$\frac{\partial\, \epsilon'\epsilon}{\partial\, \beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y \qquad (2.3.1)$$

ซึ่งจะได้ $$\hat{\beta} = b = (X'X)^{-1} X'Y \qquad (2.3.2)$$

สมการ (2.3.1) เรียกว่าสมการปกติ (normal equation) และสมการ (2.3.2) คือสมการที่ใช้ในการหาค่าประมาณของ β

การประมาณค่าโดยวิธีกำลังสองต่ำสุด จะได้ตัวประมาณเชิงเส้นที่ไม่มีความเอนเอียง และมีประสิทธิภาพสูงสุด ซึ่งเรียกว่าตัวประมาณเชิงเส้นที่ดีที่สุดที่ไม่มีความเอนเอียง (best linear unbiased estimator หรือ BLUE) และตัวประมาณโดยวิธีนี้ จะมีการแจกแจงแบบปกติ (normal distribution)

2.3.2 การประมาณค่าโดยวิธีภาวะน่าจะเป็นสูงสุด

หลักการของการหาตัวประมาณโดยวิธีนี้ ก็คือ การทำให้ฟังก์ชันความหนาแน่นร่วม (joint density function) ของตัวแปรตาม Y_i ทุก ๆ ค่า เมื่อกำหนดค่า X_{ji} มีค่ามากที่สุด ซึ่งมีขั้นตอนดังนี้

1. หาฟังก์ชันความหนาแน่นร่วมของตัวแปรตาม Y ทุกๆ ค่า เมื่อกำหนดค่า X_{ji}

เนื่องจากมีข้อสมมติเบื้องต้นว่า ความคลาดเคลื่อนเชิงสุ่ม ϵ_i กับ ϵ_j เมื่อ $i \neq j$ ไม่มีความสัมพันธ์กัน ซึ่งจะทำให้ได้ Y_i กับ Y_j เมื่อ $i \neq j$ ไม่มีความสัมพันธ์กันด้วย และจากข้อสมมติเบื้องต้นที่ว่า ความคลาดเคลื่อนเชิงสุ่ม ϵ_i มีการแจกแจงปกติ ด้วยค่าเฉลี่ย 0 ความแปรปรวนมีค่าคงที่เท่ากับ σ^2 จะได้ว่า ตัวแปร Y มีการแจกแจงปกติด้วยค่าเฉลี่ย $X\beta$ ความแปรปรวน σ^2 ดังนั้น จะได้ฟังก์ชันความหนาแน่นร่วมของตัวแปร Y ทุก ๆ ค่า เมื่อกำหนดค่า X_{ji} ซึ่งเรียกว่า ฟังก์ชันภาวะน่าจะเป็น (likelihood function) เป็น

$$L = P(Y_1, Y_2, \ldots, Y_n / X_{11}, X_{21}, \ldots, X_{q-1n})$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right]$$

2. ใส่ ln (natural logarithm ; $\log_e$) ให้กับฟังก์ชันภาวะน่าจะเป็น (บางกรณีไม่จำเป็นต้องใส่ ln) หาอนุพันธ์ของ ln L เทียบกับ β แล้วให้สมการที่ได้มีค่าเท่ากับ 0 ซึ่งจะได้สมการภาวะน่าจะเป็น (likelihood equation) เป็นดังนี้

$$\ln L = -\frac{n}{2}[\ln 2\pi + \ln \sigma^2] - \frac{1}{2\sigma^2}(Y'Y - 2\beta X'Y + \beta' X'X\beta)$$

$$\frac{\partial \ln L}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

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

จะเห็นว่า การประมาณค่าพารามิเตอร์ β โดยวิธีภาวะน่าจะเป็นสูงสุด จะให้ตัวประมาณเหมือนกับตัวประมาณจากวิธีกำลังสองต่ำสุด

2.3.3 ค่าประมาณของความแปรปรวน (Estimated variance)

จากข้อสมมติเบื้องต้นที่ว่า ความคลาดเคลื่อนเชิงสุ่ม ϵ_i มีความแปรปรวน $V(\epsilon_i)$ เท่ากับ σ^2 สามารถประมาณค่า σ^2 ได้ด้วย s^2 ซึ่งในการวิเคราะห์ความแปรปรวน (analysis of variance) จะได้ว่า s^2 ก็คือ ค่าเฉลี่ยของผลรวมกำลังสองของความคลาดเคลื่อน (mean square of error ; MSE) ดังนี้คือ

Source of variation	d.f.	Sum of squares	Mean Square
due to regression	q-1	$SSR = b'X'Y - n\bar{Y}^2$	$MSR = SSR/(q-1)$
error	n-q	$SSE = Y'Y - b'X'Y$ $= SST - SSR$	$MSE = s^2 = SSE/(n-q)$
total (about mean)	n-1	$SST = Y'Y - n\bar{Y}^2$	

2.4 การอนุมานเกี่ยวกับการถดถอย (Inferences about regression)

จากการประมาณค่าพารามิเตอร์ β ได้ค่าประมาณ คือ $\hat{\beta} = b = (X'X)^{-1}X'Y$ มีความแปรปรวน คือ $V(b) = \sigma^2 (X'X)^{-1}$ ซึ่งสามารถประมาณได้ด้วย $s^2 (X'X)^{-1}$

เมื่อ
$$s^2(X'X)^{-1} = \begin{bmatrix} s^2(b_0) & s(b_0, b_1) & \cdots & s(b_0, b_{q-1}) \\ s(b_1, b_0) & s^2(b_1) & \cdots & s(b_1, b_{q-1}) \\ \vdots & \vdots & & \vdots \\ s(b_{q-1}, b_0) & s(b_{q-1}, b_1) & \cdots & s^2(b_{q-1}) \end{bmatrix}$$

$$= \text{MSE}\,(X'X)^{-1} = \text{Cov}(b_j, b_k)$$

เรียก $\text{MSE}\,(X'X)^{-1}$ ว่า variance-covariance matrix ของค่าประมาณ b และได้ว่า การแจกแจงของตัวประมาณ b จะมีการแจกแจงแบบปกติ

สำหรับความแปรปรวนของ $\hat{Y}_o$ ซึ่งประมาณได้จาก $\hat{Y} = Xb$ เมื่อกำหนดค่า X_{jo} คือ $V(\hat{Y}_o)$

เมื่อ $V(\hat{Y}_o) = \sum_{j=0}^{q-1} X_{jo}^2 V(b_j) + 2 \sum_{j,k=0}^{q-1} X_{jo}X_{ko}\, \text{Cov}(b_j, b_k) = X_o(X'X)^{-1}X_o\, \text{MSE}$

2.4.1 ช่วงความเชื่อมั่นสำหรับพารามิเตอร์ (Confidence interval for parameters)

การประมาณค่าพารามิเตอร์ β_j ใด ๆ แบบช่วง เมื่อกำหนดช่วงความเชื่อมั่น $(1-\alpha)100$ % จะได้ค่าประมาณของ β_j มีค่าเป็น

$$b_j - t_{\alpha/2}.s(b_j) < \beta_j < b_j + t_{\alpha/2}.s(b_j) \quad ; \text{d.f.} = n-q$$

สำหรับค่าประมาณของ Y_o แบบช่วง เมื่อกำหนดช่วงความเชื่อมั่น $(1-\alpha)100$ % มีค่าเป็น

$$\hat{Y}_o - t_{\alpha/2}\sqrt{V(\hat{Y}_o)} < Y_o < \hat{Y}_o + t_{\alpha/2}\sqrt{V(\hat{Y}_o)} \quad ; \text{d.f.} = n-q$$

2.4.2 การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์

การทดสอบสมมติฐานเกี่ยวกับ β_j เพื่อทดสอบว่า ตัวแปรอิสระ X_j จะมีอิทธิพลต่อตัวแปรตาม Y หรือไม่ มีสมมติฐาน $H_o : \beta_j = 0$ และ $H_1 : \beta_j \neq 0$

จะได้ค่าทดสอบสถิติ คือ $t = b_j / s(bj)$

มีอาณาเขตวิกฤต(critical region) คือ $|t| > t_{\alpha/2,(n-q)}$

2.4.3 การทดสอบสมมติฐานเกี่ยวกับการถดถอย

การทดสอบสมมติฐานเกี่ยวกับการถดถอย ซึ่งจะทดสอบว่าตัวแปรอิสระ X_j ต่าง ๆ จะมีอิทธิพลต่อตัวแปรตาม Y หรือไม่ มีสมมติฐาน $H_o : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_{q-1} = 0$ และ H_1 : มีค่า β_j อย่างน้อย 1 ค่าที่ไม่เป็น 0

จะได้ค่าทดสอบสถิติ คือ $F = \frac{MSR}{MSE}$

อาณาเขตวิกฤต คือ $F > F_{\alpha,\{q-1,(n-q)\}}$

มีข้อควรระวังว่า การทดสอบเกี่ยวกับการถดถอยดังกล่าวข้างต้นนี้ เมื่อผลการทดสอบได้ว่าปฏิเสธ H_o มิได้หมายความว่า สมการถดถอยที่ได้จะเป็นสมการถดถอยที่เหมาะสม จะต้องมีการพิจารณาต่อไปว่าตัวแปรอิสระ X_j ใดบ้าง ควรจะอยู่ในสมการ ซึ่งจะได้กล่าวต่อไป

2.4.4 สัมประสิทธิ์ของการตัดสินใจ (Coefficient of determination)

สัมประสิทธิ์ของการตัดสินใจ เป็นค่าที่ใช้สำหรับพิจารณาว่า ตัวแปร X_j ในสมการถดถอยสามารถอธิบายตัวแปร Y หรือมีอิทธิพลต่อตัวแปร Y มากน้อยเพียงใด ทั้งนี้หมายความว่า ตัวแปร X_j ที่อยู่ในสมการได้ผ่านการทดสอบสมมติฐานแล้วว่า มีอิทธิพลต่อตัวแปรตาม Y

ถ้าให้ R^2 หมายถึงสัมประสิทธิ์ของการตัดสินใจ

จะได้ $$R^2 = \frac{\text{SS due to regression}}{\text{SS total}}$$

ถ้า R^2 มีค่าเข้าใกล้ 1 แสดงว่า ตัวแปร X_j ต่าง ๆ ในสมการ สามารถจะอธิบายตัวแปร Y ได้มาก ซึ่งหมายความว่าสมการถดถอยที่ประมาณได้เหมาะสมแล้ว แต่ถ้า R^2 มีค่าน้อย ๆ แสดงว่าตัวแปร X_j ในสมการ จะอธิบายตัวแปร Y ได้ไม่มาก ซึ่งก็หมายความว่าสมการถดถอยนั้น ยังไม่เหมาะสม จำเป็นต้องมีการตรวจสอบต่อไปว่าเป็นเพราะเหตุใด

2.5 ตัวแปรหุ่น (Dummy variable)

การวิเคราะห์การถดถอยเชิงเส้นที่ได้กล่าวมาแล้ว ค่าสังเกตของตัวแปรอิสระ X เป็นค่าวัดที่ได้จากมาตรวัดอันตรภาค (interval scale) หรือจากมาตรวัดอัตราส่วน (ratio scale) ซึ่งเป็นค่าจริง มีลักษณะเป็นค่าแบบต่อเนื่อง แต่บ่อยครั้งที่สนใจศึกษาถึงอิทธิพลของตัวแปรอิสระ X ซึ่งเป็นค่าวัดจากมาตรวัดนามบัญญัติ (nominal scale) หรือจากมาตรวัดเรียงลำดับ(ordinal scale) ซึ่งเป็นค่าวัดตามคุณภาพ มีลักษณะเป็นค่าแบบไม่ต่อเนื่อง เช่น อาชีพ ระดับการศึกษา เพศ เป็นต้น ในกรณีนี้ จะมีการนำเทคนิคของตัวแปรหุ่นมาช่วยในการวิเคราะห์การถดถอยเชิงเส้น เพื่อให้สามารถศึกษาถึงอิทธิพลของตัวแปรอิสระ ที่มีค่าวัดเป็นค่าแบบไม่ต่อเนื่องได้

เทคนิคของการใช้ตัวแปรหุ่นช่วยในการวิเคราะห์การถดถอย จะใช้ตัวแปรหุ่นที่มีค่า 0 และ 1 เท่านั้น และจำนวนตัวแปรหุ่นที่นำใส่ในสมการถดถอย จะมีจำนวนเท่ากับ m - 1 เมื่อ m หมายถึงจำนวนค่าหรือระดับ (level) ของตัวแปรอิสระที่เป็นค่าวัดแบบไม่ต่อเนื่องที่ศึกษาทั้งหมด เช่น เมื่อแบ่งอาชีพเป็น 4 กลุ่ม คือ ค้าขาย รับราชการ รับจ้าง และอื่น ๆ จะต้องใช้ตัวแปรหุ่น 0, 1 จำนวน 4 - 1 = 3 ตัว ดังนี้ คือ

X_1 แทนอาชีพค้าขาย ; $X_1 = 1$ เมื่อมีอาชีพค้าขาย
$X_1 = 0$ เมื่อมีอาชีพอื่น ๆ

X_2 แทนค่าอาชีพรับราชการ ; $X_2 = 1$ เมื่อมีอาชีพรับราชการ
$X_2 = 0$ เมื่อมีอาชีพอื่น ๆ

X_3 แทนค่าอาชีพรับจ้าง ; $X_3 = 1$ เมื่อมีอาชีพรับจ้าง
$X_3 = 0$ เมื่อมีอาชีพอื่น ๆ

และจะได้สมการถดถอยเป็น $Y = \beta_0 + \beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

การใช้ตัวแปรหุ่นช่วยในการวิเคราะห์การถดถอย ยังมีรายละเอียดที่ควรศึกษาเพิ่มเติมเพื่อให้ได้การวิเคราะห์การถดถอยที่สมบูรณ์ จะศึกษาได้จาก ธีระพงษ์ วิกิตเศรษฐ (2531), รัชนี ตียพันธ์ (2524), Draper and Smith (1966), Neter and Wasserman (1974) และ Searle (1971)

2.6 การเลือกสมการถดถอยที่ดีที่สุด

จากหัวข้อ 2.4.3 เมื่อการทดสอบสมมติฐานเกี่ยวกับการถดถอยได้ว่า ปฏิเสธ $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{q-1} = 0$ ซึ่งแสดงว่ามีค่า β_j อย่างน้อย 1 ค่า ที่ไม่เท่ากับ 0 จึง

เกิดปัญหาว่า β_j ค่าใด ที่ไม่เท่ากับ 0 หรือกล่าวอีกนัยหนึ่งก็คือ ตัวแปร X_j ตัวใดที่มีอิทธิพลต่อตัวแปร Y ซึ่งสมควรจะปรากฏอยู่ในสมการถดถอย เพื่อให้ได้สมการถดถอยที่ดีที่สุด วิธีการที่ใช้ในการพิจารณาเพื่อหาสมการถดถอยที่ดีที่สุด ได้แก่

1. วิธี all possible regressions
2. วิธี backward elimination
3. วิธี forward selection
4. วิธี stepwise regression

2.6.1 วิธี all possible regressions

การพิจารณาหาสมการถดถอยที่ดีที่สุด โดยวิธี all possible regressions ซึ่งมีตัวแปรอิสระ X_j จำนวน q ตัว ($j=0, 1, \ldots, q-1$; $X_0 = 1$) มีหลักการดังนี้

1. สร้างสมการถดถอยที่เป็นไปได้ทั้งหมด โดยเริ่มตั้งแต่สมการที่มีเฉพาะค่าคงที่หรือมีตัวแปร $X_0 = 1$ อยู่ในสมการ ไปจนถึงสมการที่มีตัวแปร X_j ครบทั้งหมดในสมการ ซึ่งจะสามารถสร้างสมการที่เป็นไปได้ทั้งหมด 2^{q-1} สมการ เช่น มีตัวแปร X_j จำนวน 4 ตัว คือ X_0, X_1, X_2 และ X_3 จะสามารถสร้างสมการถดถอย ที่เป็นไปได้ทั้งหมด = 2^3 = 8 สมการ ได้แก่

$$Y = b_0$$
$$Y = b_0 + b_1X_1$$
$$Y = b_0 + b_2X_2$$
$$Y = b_0 + b_3X_3$$
$$Y = b_0 + b_1X_1 + b_2X_2$$
$$Y = b_0 + b_1X_1 + b_3X_3$$
$$Y = b_0 + b_2X_2 + b_3X_3$$
$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

2. คำนวณค่าสัมประสิทธิ์ของการตัดสินใจ R^2 ของแต่ละสมการ แล้วพิจารณาสมการที่ให้ค่า R^2 มาก ๆ แต่ไม่จำเป็นว่า จะต้องเลือกสมการถดถอยที่ให้ค่า R^2 มากที่สุดเสมอไป และจะต้องพิจารณาสัมประสิทธิ์สหสัมพันธ์ (coefficient of correlation) ระหว่างตัวแปร Y กับตัวแปร X_j แต่ละตัวประกอบด้วย

การพิจารณาหาสมการถดถอยที่ดีที่สุดโดยวิธีนี้ จำเป็นต้องอาศัยประสบการณ์และความชำนาญของผู้ตัดสินใจเป็นอย่างมาก

2.6.2 วิธี backward elimination

หลักการของวิธี backward elimination ก็คือ การพยายามขจัดตัวแปรอิสระ X_j ที่ไม่มีนัยสำคัญออกทีละตัว ซึ่งมีขั้นตอนดังนี้

1. สร้างสมการถดถอยที่มีตัวแปรอิสระ X_j ทุกตัวในสมการ และทดสอบ $H_o : \beta_1 = \beta_1 = \ldots = \beta_{q-1} = 0$ เมื่อผลการทดสอบได้ว่าปฏิเสธ H_o จะพิจารณาตัดตัวแปร X_j ออกจากสมการ

2. คำนวณค่า partial F-test ของแต่ละตัวแปร X_j ทุกๆ ตัว ในลักษณะที่ตัวแปร X_j นั้น ๆ เข้ามาในสมการเป็นตัวสุดท้าย

3. พิจารณาค่า partial F-test ของตัวแปร X_j ที่มีค่าต่ำสุด ทำการทดสอบค่า β_j ของตัวแปรนั้นว่ามีค่าเท่ากับ 0 หรือไม่ ($H_o : \beta_j = 0$) ถ้าผลการทดสอบได้ว่าปฏิเสธ H_o ก็แสดงว่าตัวแปรตามขั้นตอนที่ 1 ทุกตัว ควรจะอยู่ในสมการถดถอยทั้งหมด แต่ถ้าผลทดสอบได้ว่า ยอมรับ H_o ก็จะตัดตัวแปร X_j ที่มีค่า partial F-test ต่ำสุดนี้ออกจากสมการ

4. สร้างสมการถดถอยใหม่ โดยไม่ให้มีตัวแปรที่ถูกตัดทิ้ง จากขั้นตอนที่ 3 อยู่ในสมการ แล้วเริ่มดำเนินการเหมือนขั้นตอนที่ 1 - 3 จนได้สมการถดถอยที่ต้องการ

2.6.3 วิธี forward selection

หลักการที่ใช้ในการเลือกสมการถดถอยที่ดีที่สุด โดยวิธี forward selection ก็คือ พยายามเลือกตัวแปรอิสระ X_j ที่มีความสัมพันธ์กับตัวแปร Y มากที่สุด เข้าไปในสมการก่อน โดยจะพิจารณาจากสัมประสิทธิ์สหสัมพันธ์เชิงส่วน (partial correlation coefficient) มีขั้นตอนดังนี้

1. เลือกตัวแปร X_j ที่มีความสัมพันธ์กับตัวแปร Y มากที่สุด เข้าไปในสมการถดถอยก่อน โดยการทดสอบความสัมพันธ์ก่อนว่ามีนัยสำคัญ แล้วทดสอบ $H_o : \beta_j = 0$ ถ้าผลการทดสอบได้ว่า ปฏิเสธ H_o จะพิจารณาเพิ่มตัวแปร X_j เข้าในสมการอีก ตามขั้นตอนที่ 2

2. คำนวณสัมประสิทธิ์สหสัมพันธ์เชิงส่วน ระหว่างตัวแปร X_j ที่เหลือ กับตัวแปร Y เมื่อมีตัวแปร X_j จากขั้นตอนที่ 1 ในสมการแล้ว ทุก ๆ ตัว นำตัวแปร X_j ที่มีค่าสัมประสิทธิ์สหสัมพันธ์เชิงส่วนมากที่สุดเข้าในสมการ โดยต้องทดสอบความสัมพันธ์ก่อนว่ามีนัยสำคัญ แล้วทดสอบ $H_o : \beta_j$ (ของตัวแปรใหม่) $= 0$ ถ้าผลการทดสอบได้ว่ายอมรับ H_o ก็แสดงว่าตัวแปรตามขั้นตอนที่ 2 นี้ จะไม่ถูกนำเข้าไปใส่ในสมการ ซึ่งก็จะได้สมการถดถอยมีเฉพาะตัวแปร X_j ตามขั้นตอนที่ 1 อยู่ในสมการ เท่านั้น แต่ถ้าผลการทดสอบได้ว่าปฏิเสธ H_o ก็จะพิจารณานำตัวแปร X_j ตัวใหม่ ใส่เข้าไปในสมการอีก โดยการทำซ้ำขั้นตอนที่ 2 ไปเรื่อยๆ จนได้สมการถดถอยตามที่ต้องการ

2.6.4 วิธี stepwise regression

การเลือกสมการถดถอยที่ดีที่สุด โดยวิธี stepwise regression จะเป็นการผสมผสานกันระหว่างวิธี backward elimination กับวิธี forward selection แล้วปรับปรุงให้ดีขึ้น มีขั้นตอนพอสรุปได้ดังนี้

1. นำตัวแปร X_j ที่ความสัมพันธ์กับตัวแปร Y มากที่สุด เข้าในสมการก่อน แล้วทดสอบ $H_o : \beta_j = 0$

2. เมื่อการทดสอบ $H_o : \beta_j = 0$ ในขั้นตอนที่ 1 พบว่า ปฏิเสธ H_o ก็จะพิจารณาตัวแปร X_j ที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์เชิงส่วน มีค่ามากที่สุด นำเข้าสมการ แล้วทดสอบ H_o : all $\beta_j = 0$ และ แต่ละ $H_o : \beta_j = 0$ โดยการใช้ partial F-Test ถ้าผลการทดสอบจากการใช้ partial F-Test ของ $H_o : \beta_j = 0$ ใด ได้ว่ายอมรับ H_o ก็จะนำตัวแปร X_j นั้นออกจากสมการ แต่ถ้าผลการทดสอบได้ว่าปฏิเสธ H_o ก็จะพิจารณาเพิ่มตัวแปร X_j ที่เหลือเข้าไปใหม่อีก

3. ดำเนินการตามขั้นตอนที่ 2 จนได้สมการถดถอยที่ต้องการ

2.7 ความเหมาะสมของโมเดล

หลังจากเลือกโมเดลการถดถอยเชิงเส้น ตลอดจนประมาณค่าพารามิเตอร์ต่าง ๆ ได้แล้ว ก่อนนำสมการถดถอยเชิงเส้นที่ประมาณได้ไปใช้ต่อไป ควรจะแน่ใจว่าโมเดลที่ได้นั้นมีความเหมาะสมเพียงพอแล้ว มีบ่อยครั้งที่เลือกโมเดลการถดถอยเชิงเส้นและทดสอบสมมติฐานแล้วพบว่า สมการถดถอยเชิงเส้นที่ประมาณได้ให้ค่าสัมประสิทธิ์ของการตัดสินใจต่ำเกินไป ซึ่งเป็นไปได้ว่าโมเดลที่ได้นั้นยังไม่เหมาะสม ในหัวข้อนี้จะเป็นการพิจารณาหาสาเหตุของความไม่เหมาะสมของโมเดล จากค่าความคลาดเคลื่อน

2.7.1 ความคลาดเคลื่อนของตัวอย่าง

ถ้า Y_i คือค่าสังเกตของตัวแปรตาม ณ ค่า X_{ji} ที่กำหนด และ $\hat{Y}_i$ คือค่าที่ประมาณได้จากสมการถดถอยเชิงเส้น $\hat{Y}_i = \sum_{j=0}^{q-1} b_j X_{ji}$ จะได้ความคลาดเคลื่อน คือ e_i เป็นดังนี้

$$e_i = Y_i - \hat{Y}_i$$

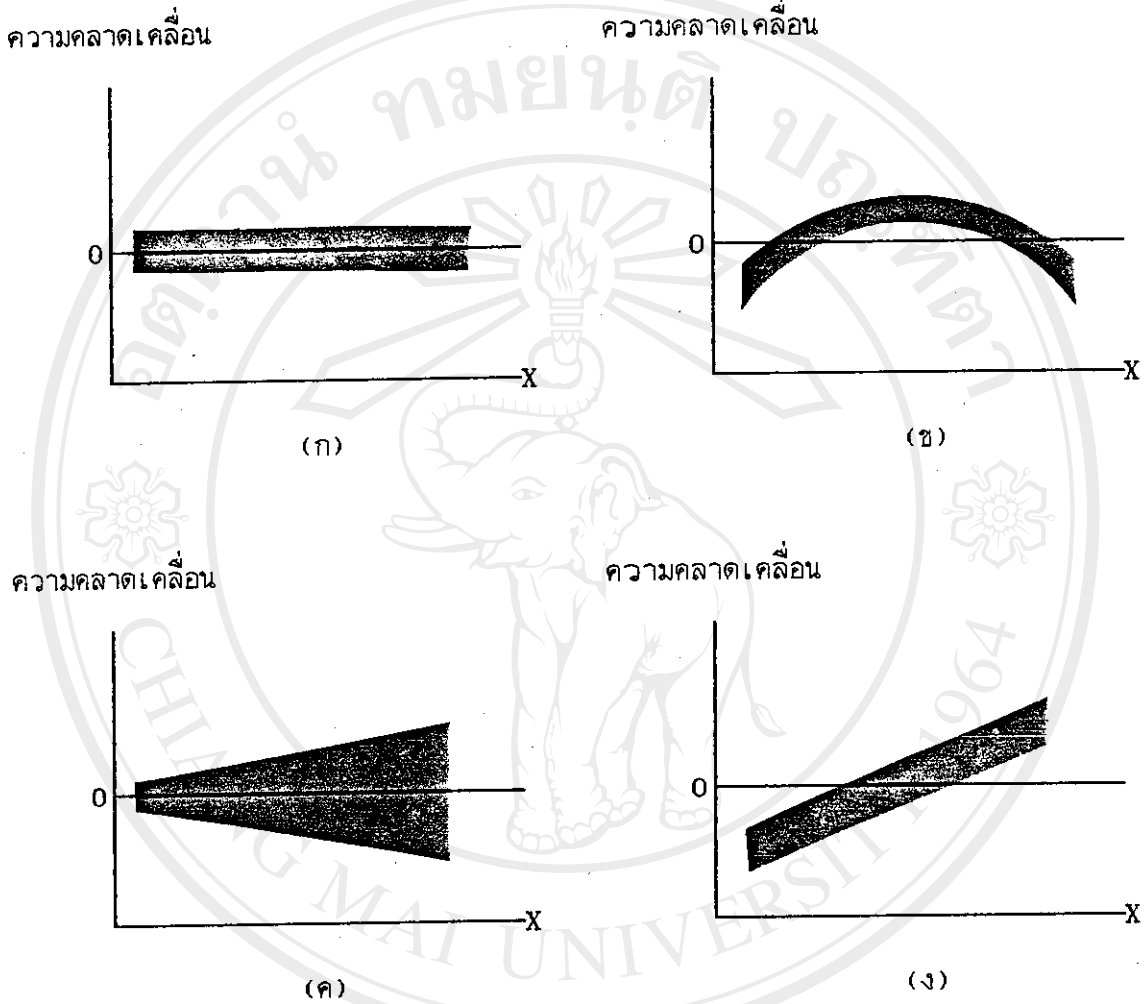
และจากข้อสมมติเบื้องต้นของความคลาดเคลื่อนเชิงสุ่ม ϵ_i ว่า เป็นตัวแปรเชิงสุ่มที่มีการแจกแจงแบบปกติที่เป็นอิสระกัน ด้วยค่าเฉลี่ย 0 ความแปรปรวนคงที่เท่ากับ σ^2 ดังนั้น ความคลาดเคลื่อน e_i ควรจะมีค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ $MSE = s^2$ และจะได้ความคลาดเคลื่อนมาตรฐาน(standardized residual) เป็น $\frac{e_i}{\sqrt{MSE}}$

จะใช้ความคลาดเคลื่อน และความคลาดเคลื่อนมาตรฐาน ในการพิจารณาหาสาเหตุของความไม่เหมาะสมของโมเดล โดยดูจากกราฟของความคลาดเคลื่อนหรือความคลาดเคลื่อนมาตรฐาน ในรูปแบบต่าง ๆ ซึ่งจะบ่งชี้ถึงสาเหตุของความไม่เหมาะสมของโมเดล ในประเด็นที่เกี่ยวข้องกับข้อสมมติเบื้องต้นได้ สำหรับการทดสอบเชิงสถิติเกี่ยวกับความคลาดเคลื่อนมีรายละเอียดมาก จะกล่าวเฉพาะวิธีการที่ใช้ในการทดสอบเท่านั้น รายละเอียดสามารถศึกษาเพิ่มเติมได้จาก วัชรี พฤกษิกานนท์ (2528), Draper and Smith (1966) และ Neter and Wasserman (1974)

2.7.2 การวิเคราะห์ความคลาดเคลื่อนโดยกราฟ

กราฟของความคลาดเคลื่อนสามารถทำได้หลายรูปแบบ โดยปกติจะใช้กราฟแสดงความสัมพันธ์ระหว่างความคลาดเคลื่อน e_i ในแนวตั้ง กับค่าของตัวแปร Y_i ที่ประมาณจากสมการถดถอยเชิงเส้น หรือกับตัวแปรอิสระ $X_{j,i}$ ในแนวนอน จากกราฟ จะดูการกระจายของความคลาดเคลื่อน รอบเส้นตรงที่ขนานกับแนวนอน ณ ค่า $e_i = 0$ ว่า มีลักษณะเป็นอย่างไร

ถ้ากราฟที่ได้มีลักษณะคล้ายรูปที่ 2.7.1 (ก) แสดงว่าสมการถดถอยที่ใช้มีลักษณะเป็นเชิงเส้น ความคลาดเคลื่อนเป็นอิสระกัน และมีความแปรปรวนคงที่ ซึ่งมีความหมายว่าการประมาณสมการถดถอยด้วยสมการเชิงเส้นนั้นถูกต้องแล้ว การที่สมการไม่เหมาะสมอาจเนื่องมาจากสาเหตุอื่น เช่น อาจเนื่องจากในสมการขาดตัวแปรอิสระที่สำคัญ ซึ่งมีอิทธิพลต่อตัวแปรตาม เป็นต้น



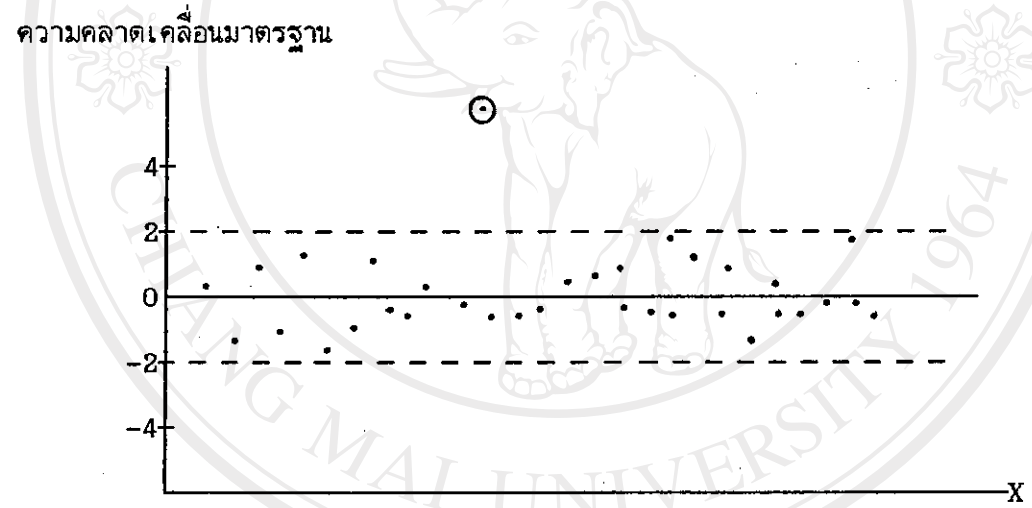
รูปที่ 2.7.1

ถ้ากราฟที่ได้มีลักษณะคล้ายรูปที่ 2.7.1 (ข) แสดงว่าการใช้สมการถดถอยเป็นเส้นตรงจะไม่ถูกต้อง ควรมีการเพิ่มเทอมกำลังสองของตัวแปรอิสระบางตัวให้กับสมการ

ถ้ากราฟที่ได้มีลักษณะคล้ายรูปที่ 2.7.1 (ค) แสดงว่าความคลาดเคลื่อนมีค่าความแปรปรวนไม่คงที่ จำเป็นต้องมีการแปลงค่าสังเกต Y_i หรือใช้การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองต่ำสุดแบบถ่วงน้ำหนัก เพื่อทำให้ความคลาดเคลื่อนมีค่าความแปรปรวนคงที่

ถ้ากราฟที่ได้มีลักษณะคล้ายรูปที่ 2.7.1 (ง) แสดงว่าความคลาดเคลื่อนมีความสัมพันธ์กัน

นอกจากนี้แล้ว อาจพิจารณาจากกราฟแสดงความสัมพันธ์ระหว่างความคลาดเคลื่อนมาตรฐาน กับค่าของตัวแปรตาม Y_i ที่ประมาณได้ หรือกับตัวแปร X_{ji} ซึ่งนอกจากจะบอกให้ทราบว่า ความคลาดเคลื่อนมีการแจกแจงแบบปกติหรือไม่ โดยดูจากการกระจายของความคลาดเคลื่อนมาตรฐาน รอบแนวนอน ณ ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ 0 ± 2 แล้วยังใช้พิจารณาว่าจะมีค่าสังเกตที่ผิดปกติรวมอยู่ในข้อมูลหรือไม่ โดยดูจากความคลาดเคลื่อนมาตรฐานที่อยู่นอกขอบเขต 0 ± 2 ตามรูปที่ 2.7.2



รูปที่ 2.7.2

2.7.3 การทดสอบเกี่ยวกับความคลาดเคลื่อน

การทดสอบเชิงสถิติที่ใช้ทดสอบเกี่ยวกับข้อสมมติเบื้องต้น ของการวิเคราะห์การถดถอยเชิงเส้น พอสรุปได้ดังนี้

1. การทดสอบว่าความคลาดเคลื่อน มีความสัมพันธ์กันหรือไม่ ใช้การทดสอบ Durbin-Watson

2. การทดสอบว่าความคลาดเคลื่อน มีความแปรปรวนคงที่หรือไม่ อาจใช้วิธีง่าย ๆ โดยการแบ่งข้อมูลออกเป็น 2 ส่วน คำนวณความแปรปรวนของความคลาดเคลื่อนในแต่ละส่วน แล้วทดสอบการเท่ากันของความแปรปรวนด้วยการทดสอบแบบเอฟ (F-test)

3. การทดสอบความคลาดเคลื่อนว่า มีการแจกแจงปกติหรือไม่ ใช้การทดสอบภาวะสารูปดี ซึ่งอาจใช้การทดสอบแบบไคสแควร์ (Chi-square test) หรือการทดสอบ Komogorov-Smirnov

4. การทดสอบว่าสมการถดถอยเป็นเชิงเส้นหรือไม่ จะใช้การทดสอบแบบเอฟ

2.7.4 แนวทางแก้ไขเมื่อเกิดปัญหาโมเดลไม่เหมาะสม

เมื่อเกิดปัญหาโมเดลไม่เหมาะสม อาจเลิกใช้โมเดลเดิม แล้วหาโมเดลใหม่ที่เหมาะสมมาใช้แทน ซึ่งกรณีนี้อาจประสบกับปัญหาของการประมาณค่าพารามิเตอร์ ว่ามีความยุ่งยากมากขึ้นก็ได้ โดยทั่วไป นิยมแก้ไขปัญหาด้วยการแปลงข้อมูล ซึ่งอาจเป็นการแปลงข้อมูลของตัวแปรตาม แปลงข้อมูลของตัวแปรอิสระ หรือแปลงข้อมูลทั้งสมการ แล้วแต่กรณี เพื่อให้การวิเคราะห์การถดถอยเป็นไปอย่างเหมาะสม เช่น

เพื่อให้ได้ค่าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ อาจแปลงข้อมูลของตัวแปรตาม เป็น

$$Y' = \sqrt{Y}$$

หรือ $$Y' = \sin^{-1}\sqrt{Y}$$

หรือ $$Y' = \ln(Y)$$

หรือใช้การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองต่ำสุดแบบถ่วงน้ำหนัก

และในกรณีที่มีตัวแปรอิสระบางตัวไม่เหมาะสม อาจแปลงค่าของตัวแปรอิสระนั้น ๆ ก่อนนำเข้าสมการ เป็น $X' = \log X$ เป็นต้น