บทที่ 4

การตรวจสอบความเหมาะสมของสมการถดถอย

เป็นที่ทราบแล้วว่าข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยประกอบด้วย ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระมีลักษณะเชิงเส้นตรง ความคลาดเคลื่อน ϵ_i มีค่าเฉลี่ยเป็น 0 ความแปรปรวนคงที่เท่ากับ σ^2 และมีการแจกแจงแบบปกติ โดย ϵ_i และ ϵ_j $(i \neq j)$ ต้องเป็นอิสระกัน ซึ่งการอนุมานทางสถิติไม่ว่าจะเป็นการทดสอบสมมติฐานหรือการประมาณช่วงความเชื่อมั่น ต้องอาศัยข้อกำหนดเกี่ยวกับการแจกแจงแบบปกติของความคลาดเคลื่อน โดยก่อนที่จะรับเอาโมเดลภายใต้การพิจารณาไปใช้ ควรที่จะมีการตรวจสอบความเหมาะสมของโมเดลก่อน ซึ่งการตรวจสอบดังกล่าวไม่สามารถทำได้ด้วยสถิติทั่วไป เช่น สถิติทดสอบ t หรือ F ในบทนี้จะกล่าวถึงวิธีการตรวจสอบข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย ซึ่งสามารถประยุกต์ใช้ได้กับการวิเคราะห์ การถดถอยเชิงตรงเส้นอย่างง่ายและแบบพหุ รวมทั้งระบุวิธีแก้ไขเมื่อข้อมูลมีลักษณะไม่สอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์

4.1 การวิเคราะห์ความคลาดเคลื่อนตัวอย่าง (Residual Analysis)

ความคลาดเคลื่อนตัวอย่างเกิดจากความแตกต่างระหว่างค่าจริงและค่าที่อยู่บนเส้นถดถอย สามารถแสดงได้ดังนี้

$$e_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots, n \tag{4.1}$$

เมื่อ

Y_i แทน ค่าจริงที่สังเกตได้

$\hat{Y}_i$ แทน ค่าพยากรณ์ที่ได้จากสมการถดถอย

โดยความคลาดเคลื่อนตัวอย่างเป็นค่าที่ใช้อธิบายความผันแปรที่ไม่สามารถอธิบายได้ด้วยสมการถดถอย ในที่นี้จะถือว่าชุดข้อมูลที่พิจารณาประกอบด้วยความคลาดเคลื่อนที่สังเกตได้ หากความคลาดเคลื่อนที่ได้จาก

โมเดลถดถอยมีลักษณะเบี่ยงเบนไปจากข้อกำหนดของการวิเคราะห์ ลักษณะที่เบี่ยงเบนดังกล่าวควรจะปรากฏในความคลาดเคลื่อนตัวอย่างด้วย ดังนั้นการวิเคราะห์ความคลาดเคลื่อนตัวอย่างจึงจัดเป็นวิธีที่มีประสิทธิภาพในการตรวจสอบข้อบกพร่องหลายประการของโมเดล

จากบทที่ 2 ทราบแล้วว่าความคลาดเคลื่อนตัวอย่างมีค่าเฉลี่ยเป็น 0 นั่นคือ $\bar{e} = 0$ และค่าประมาณของความแปรปรวนเป็น

$$S^2 = MSE = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-k-1} \tag{4.2}$$

โดยความคลาดเคลื่อนตัวอย่างไม่เป็นอิสระต่อกัน แต่จะมีผลกระทบต่อการตรวจสอบความเหมาะสมของโมเดลน้อยมากหากขนาดตัวอย่าง n ไม่เล็กจนเกินไป

การตรวจสอบข้อกำหนดเบื้องต้นของการวิเคราะห์การถดถอย นอกจากจะตรวจสอบจากความคลาดเคลื่อนตัวอย่างโดยตรงแล้ว ยังอาจตรวจสอบโดยใช้ความคลาดเคลื่อนตัวอย่างในรูปแบบอื่นได้อีก ดังนี้

- **Standardized residuals** แทนด้วย d_i มีสูตรดังนี้

$$d_i = \frac{e_i}{MSE}, \qquad i = 1, 2, \ldots, n \tag{4.3}$$

ซึ่ง d_i มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ 1 โดยประมาณ จะเห็นได้ว่าสมการ (4.3) เป็นการปรับค่าความคลาดเคลื่อนตัวอย่างด้วยการหารด้วยค่าประมาณของส่วนเบี่ยงเบนมาตรฐานเฉลี่ย

- **Studentized residuals** แทนด้วย r_i มีสูตรดังนี้

$$r_i = \frac{e_i}{\sqrt{MSE\,(1-h_i)}} = \frac{e_i}{S\sqrt{1-h_i}}, \qquad i = 1, 2, \ldots, n \tag{4.4}$$

เมื่อ h_{ii} แทน ค่า *Leverage* หรือ *Hat value*

จะเห็นได้ว่าสมการ (4.4) เป็นการปรับค่าความคลาดเคลื่อนตัวอย่างด้วยการหารด้วยค่าความคลาดเคลื่อนมาตรฐานที่แท้จริงของความคลาดเคลื่อนตัวอย่าง โดยพิจารณาจากการวิเคราะห์การถดถอยเชิงเส้นตรง

อย่างง่ายต่อไปนี้

$$\begin{aligned} V(e_i) &= V(Y_i - \hat{Y}_i) \\ &= V(Y_i) + V(\hat{Y}_i) - 2Cov(Y_i, \hat{Y}_i) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right] - 2Cov(Y_i, \hat{Y}_i) \end{aligned}$$

เนื่องจาก

$$\begin{aligned} \hat{Y}_i &= b_0 + b_1 X_i \\ &= (\bar{Y} - b_1 \bar{X}) + \frac{S_{xy}}{S_{xx}} X_i \\ &= \left(\bar{Y} - \frac{S_{xy}}{S_{xx}} \bar{X}\right) + \frac{S_{xy}}{S_{xx}} X_i \\ &= \bar{Y} + \frac{S_{xy}}{S_{xx}} (X_i - \bar{X}) \end{aligned}$$

จะได้ว่า

$$\begin{aligned} Cov(Y_i, \hat{Y}_i) &= Cov\left(Y_i, \bar{Y} + \frac{S_{xy}}{S_{xx}}(X_i - \bar{X})\right) \\ &= Cov\left(Y_i, \bar{Y}\right) + Cov\left(Y_i, \frac{S_{xy}}{S_{xx}}(X_i - \bar{X})\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right) \end{aligned}$$

ดังนั้น ความแปรปรวนของความคลาดเคลื่อนตัวอย่างตัวที่ i คือ

$$\begin{aligned} V(e_i) &= V(Y_i) + V(\hat{Y}_i) - 2Cov(Y_i, \hat{Y}_i) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right) - 2\sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right) \\ &= \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right)\right] \\ &= \sigma^2 (1 - h_i) \end{aligned}$$

เมื่อ $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}$

เนื่องจากไม่ทราบค่า σ^2 ดังนั้นจะประมาณด้วย S^2 หรือ MSE และได้ค่าประมาณความแปรปรวนของ

ความคลาดเคลื่อนตัวอย่างเป็น

$$\hat{V}(e_i) = MSE(1 - h_i) = S^2(1 - h_i) \tag{4.5}$$

ทำนองเดียวกันสำหรับการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุจะได้ว่า

$$r_i = \frac{e_i}{\sqrt{MSE\,(1 - h_{ii})}} = \frac{e_i}{S\sqrt{1 - h_{ii}}}, \qquad i = 1, 2, \ldots, n \tag{4.6}$$

เมื่อ h_{ii} เป็นค่าที่อยู่บนแนวทะแยงมุมในตำแหน่งที่ $(i,\ i)$ ของ Hat matrix $\mathbf{H}$
โดยที่ $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

กรณีที่ตัวอย่างมีขนาดเล็ก ค่า Studentized residuals จะเหมาะสมมากกว่าค่า Standardized residuals เพราะหากความแปรปรวนของความคลาดเคลื่อนมีค่าแตกต่างกันมาก จะส่งผลกระทบต่อ Standardized residuals ค่อนข้างมาก ทั้งนี้เนื่องจาก Standardized residuals มีการปรับค่าความคลาดเคลื่อนตัวอย่างโดยหารด้วย MSE เพียงอย่างเดียว แต่ถ้าตัวอย่างมีขนาดใหญ่ ทั้ง Standardized residuals และ Studentized residuals จะมีค่าใกล้เคียงกัน

4.2 กราฟของความคลาดเคลื่อน (Residual Plots)

การตรวจสอบความเหมาะสมของโมเดลจากความคลาดเคลื่อนตัวอย่างจะพิจารณาจากกราฟของความคลาดเคลื่อนซึ่งควรที่จะตรวจสอบทุกครั้งควบคู่ไปกับการสร้างสมการถดถอย

4.2.1 Normal Probability Plot

ถึงแม้ว่าการแจกแจงของความคลาดเคลื่อนที่เบี่ยงเบนไปจากการแจกแจงแบบปกติเล็กน้อย อาจไม่มีผลกระทบต่อโมเดลมากนัก แต่โดยรวมแล้วอาจมีผลต่อการวิเคราะห์ที่รุนแรงได้ เนื่องจากสถิติทดสอบ t หรือ F ที่ใช้ในการทดสอบสมมติฐาน รวมทั้งการสร้างช่วงความเชื่อมั่นและช่วงแห่งการพยากรณ์ต้องอาศัยข้อกำหนดเกี่ยวกับการแจกแจงแบบปกติทั้งสิ้น นอกจากนี้ถ้าความคลาดเคลื่อนมาจากการแจกแจงที่มีหางหนากว่าการแจกแจงแบบปกติแล้ว มักจะมีค่าสังเกตที่ผิดปกติ หรือที่เรียกว่า *Outliers* ปะปนอยู่ด้วย ซึ่งอาจส่งผลให้ตัวประมาณกำลังสองน้อยที่สุดถูกดึงเข้าหาค่าที่ผิดปกติเหล่านั้น ดังนั้นควรใช้วิธีประมาณพารามิเตอร์วิธีอื่นเมื่อความคลาดเคลื่อนมีการแจกแจงดังกล่าว

การสร้าง Normal probability plot เป็นวิธีการหนึ่งที่ใช้ในการตรวจสอบการแจกแจงแบบปกติของความคลาดเคลื่อน ทำได้โดยพล็อตค่าความคลาดเคลื่อนที่เรียงลำดับแล้ว $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$ กับความน่าจะ-

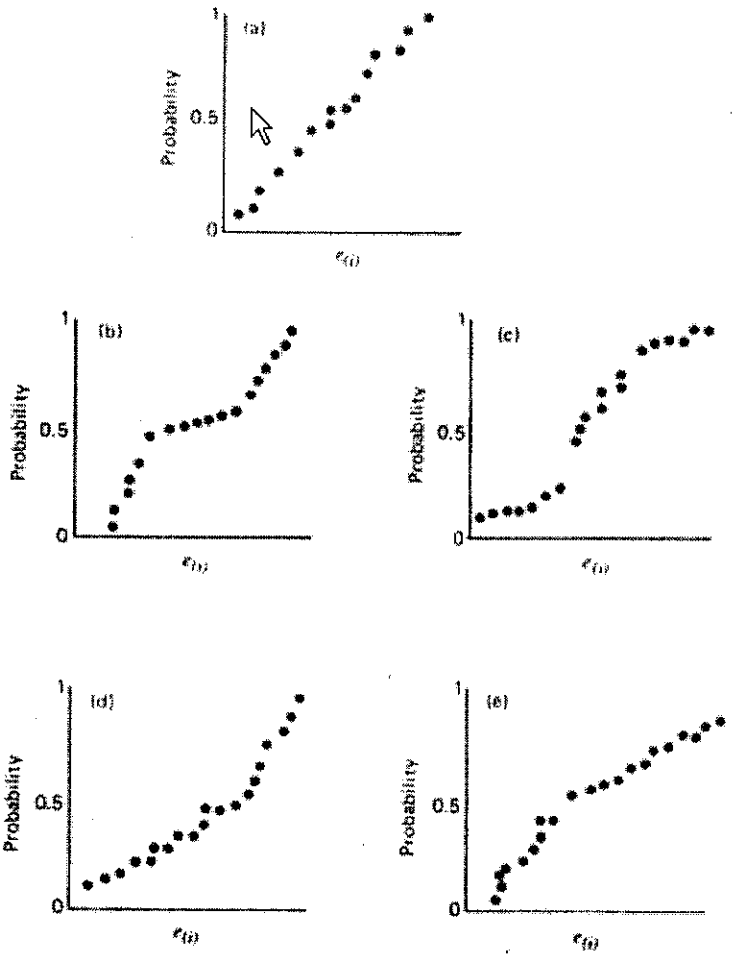
เป็นสะสม (Cumulative probability) นั่นคือ

$$P_i = \left(\frac{i - \frac{1}{2}}{n}\right), \qquad i = 1, 2, \ldots, n \tag{4.7}$$

บนกระดาษ Normal probability หรือพล็อต $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$ กับค่าคาดหวังแบบปกติของมัน (Expected normal value) ซึ่งบางครั้งเรียกว่า *Rankits* และสามารถคำนวณได้จากสูตรต่อไปนี้

$$\text{ค่าคาดหวัง} = \Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right) \cdot \sqrt{MSE}, \qquad i = 1, 2, \ldots, n \tag{4.8}$$

เมื่อ Φ แทน การแจกแจงสะสมแบบปกติมาตรฐาน (Standard normal cumulative distribution)



รูปที่ 4.1: Normal probability plot สำหรับข้อมูลที่มี (a) การแจกแจงปกติมาตรฐาน (b) การแจกแจงที่มีหางหนา (c) การแจกแจงที่มีหางบาง (d) การแจกแจงเบ้ขวา (e) การแจกแจงเบ้ซ้าย (จาก Montgomery และ Peck (1992))

รูปที่ 4.1 แสดงลักษณะต่าง ๆ ของ Normal probability plots ถ้าความคลาดเคลื่อนมีการแจกแจงแบบปกติแล้ว จุดเหล่านี้ควรตกใกล้เคียงกับเส้นตรงโดยประมาณ ดังแสดงในรูปที่ 4.1 (a) ส่วนรูปที่ 4.1 (b)-(e) แสดง

ถึงปัญหาที่เกิดขึ้น เนื่องจากข้อมูลมีการแจกแจงเบี่ยงเบนไปจากการแจกแจงแบบปกติ ในบางครั้งถึงแม้ว่าตัวอย่างจะถูกสุ่มมาจากประชากร ที่มีการแจกแจงแบบปกติ แต่กราฟของความคลาดเคลื่อนอาจไม่มีลักษณะเป็นเส้นตรงก็ได้ ดังนั้นการแปลผลของ Normal probability plots จึงต้องอาศัยประสบการณ์และความชำนาญในการพิจารณาว่าลักษณะที่เบี่ยงเบนไปจากเส้นตรงเพียงไรที่ยังสามารถยอมรับได้ นอกจากนี้ Normal probability plots อาจแสดงให้เห็นถึงข้อมูลที่มีค่ามากผิดปกติได้อีกด้วย

การแจกแจงแบบปกติยังสามารถตรวจสอบได้อีกวิธีโดยการสร้างฮิสโตแกรมของความคลาดเคลื่อน แต่อย่างไรก็ตามถ้าข้อมูลมีจำนวนน้อยจนเกินไป จะทำให้ไม่เห็นรูปร่างของการแจกแจงที่ชัดเจน นอกจากนี้ทั้ง Standardized residuals (d_i) และ Studentized residuals (r_i) ยังสามารถใช้ตรวจสอบการแจกแจงแบบปกติได้เช่นเดียวกับค่าความคลาดเคลื่อนปกติ (e_i) หากข้อมูลมีการแจกแจงแบบปกติแล้ว ประมาณ 68% ของ Standardized residuals ควรตกอยู่ระหว่าง -1 ถึง $+1$ และประมาณ 95% ตกอยู่ระหว่าง -2 ถึง $+2$ หากร้อยละของข้อมูลแตกต่างไปจากขอบเขตที่กำหนดมาก ๆ อาจชี้ว่าข้อมูลไม่มีการแจกแจงแบบปกติ นอกจากนี้ถ้าตัวอย่างมีขนาดเล็ก อาจแทนค่า ± 1 และ ± 2 ด้วยค่าที่ได้จากการแจกแจงแบบ t_{n-2} ที่สอดคล้องกัน ซึ่งจะเห็นได้ว่า Standardized residuals สามารถใช้ตรวจสอบข้อมูลที่ผิดปกติได้เช่นกัน ด้วยเหตุนี้นักวิเคราะห์บางคนจึงนิยมสร้าง Normal probability plots โดยใช้ d_i หรือ r_i แทน e_i

ตัวอย่างที่ 4.1 ในการศึกษาความสัมพันธ์ระหว่างยอดขาย (หน่วย: หมื่นบาท) และค่าใช้จ่ายในการโฆษณา (หน่วย: หมื่นบาท) ของบริษัท 16 แห่ง คำนวณสมการถดถอยได้เป็น $\hat{Y}_i = 61.184 + 5.0103X_i$ และ $MSE = 22.5287$ แล้วหาค่า ความคลาดเคลื่อนตัวอย่าง (e_i) ได้ดังนี้

บริษัท	ค่าใช้จ่ายในการโฆษณา (X_i)	ยอดขาย (Y_i)	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
1	5	89	86.2353	2.7647
2	6	87	91.2456	-4.2456
3	7	98	96.2559	1.7441
4	8	110	101.2662	8.7338
5	9	103	106.2765	-3.2765
6	10	114	111.2868	2.7132
7	11	116	116.2971	-0.2971
8	12	110	121.3074	-11.3074
9	13	126	126.3176	-0.3176
10	14	130	131.3279	-1.3279
11	15	140	136.3382	3.6618
12	16	138	141.3485	-3.3485
13	17	145	146.3588	-1.3588
14	18	153	151.3691	1.6309
15	19	155	156.3794	-1.3794
16	20	167	161.3897	5.6103

ตารางที่ 4.1: ข้อมูลยอดขายและค่าใช้จ่ายในการโฆษณา

จงสร้าง Normal probability plot ของความคลาดเคลื่อนตัวอย่างที่ได้จากสมการถดถอยเชิงเส้นตรงอย่างง่าย แสดงความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณา

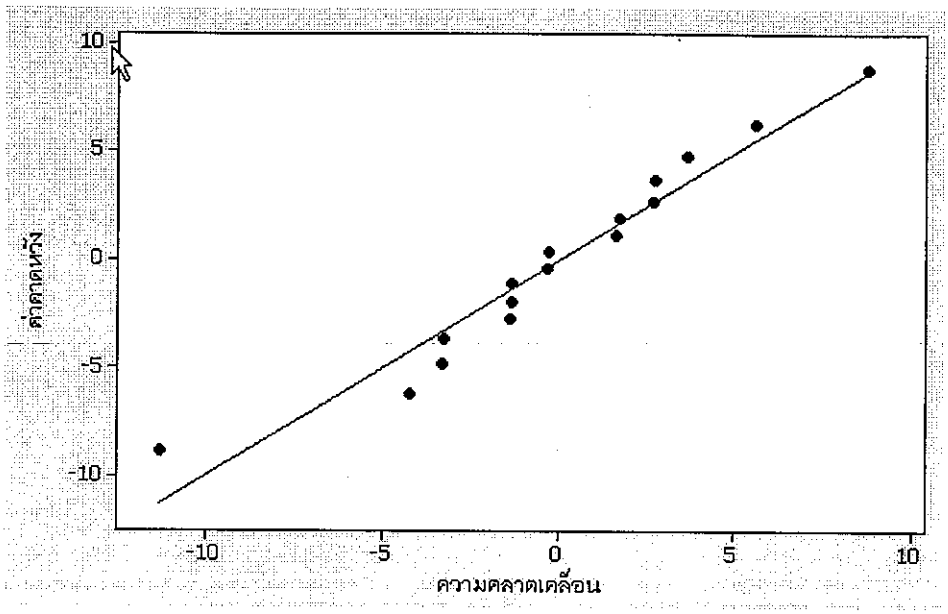
วิธีทำ เรียงลำดับความคลาดเคลื่อนตัวอย่างจากน้อยไปหามาก แล้วคำนวณค่าคาดหวังแบบปกติได้ดังนี้

$$\text{ค่าคาดหวัง} = \Phi^{-1}\left(\frac{i-\frac{1}{2}}{16}\right)\cdot\sqrt{22.5287}$$

ลำดับที่	$e_{(i)}$	$\frac{i-\frac{1}{2}}{16}$	$\Phi^{-1}\left(\frac{i-\frac{1}{2}}{16}\right)$	ค่าคาดหวัง
1	-11.3074	0.0313	-1.8627	-8.8413
2	-4.2456	0.0938	-1.3180	-6.2559
3	-3.3485	0.1563	-1.0100	-4.7939
4	-3.2765	0.2188	-0.7764	-3.6852
5	-1.3794	0.2813	-0.5791	-2.7488
6	-1.3588	0.3438	-0.4023	-1.9093
7	-1.3279	0.4063	-0.2372	-1.1259
8	-0.3176	0.4688	-0.0784	-0.3722
9	-0.2971	0.5313	0.0784	0.3722
10	1.6309	0.5938	0.2372	1.1259
11	1.7441	0.6563	0.4023	1.9093
12	2.7132	0.7188	0.5791	2.7488
13	2.7647	0.7813	0.7764	3.6852
14	3.6618	0.8438	1.0100	4.7939
15	5.6103	0.9063	1.3180	6.2559
16	8.7338	0.9688	1.8627	8.8413

ตารางที่ 4.2: การหาค่าคาดหวังแบบปกติของความคลาดเคลื่อนตัวอย่าง

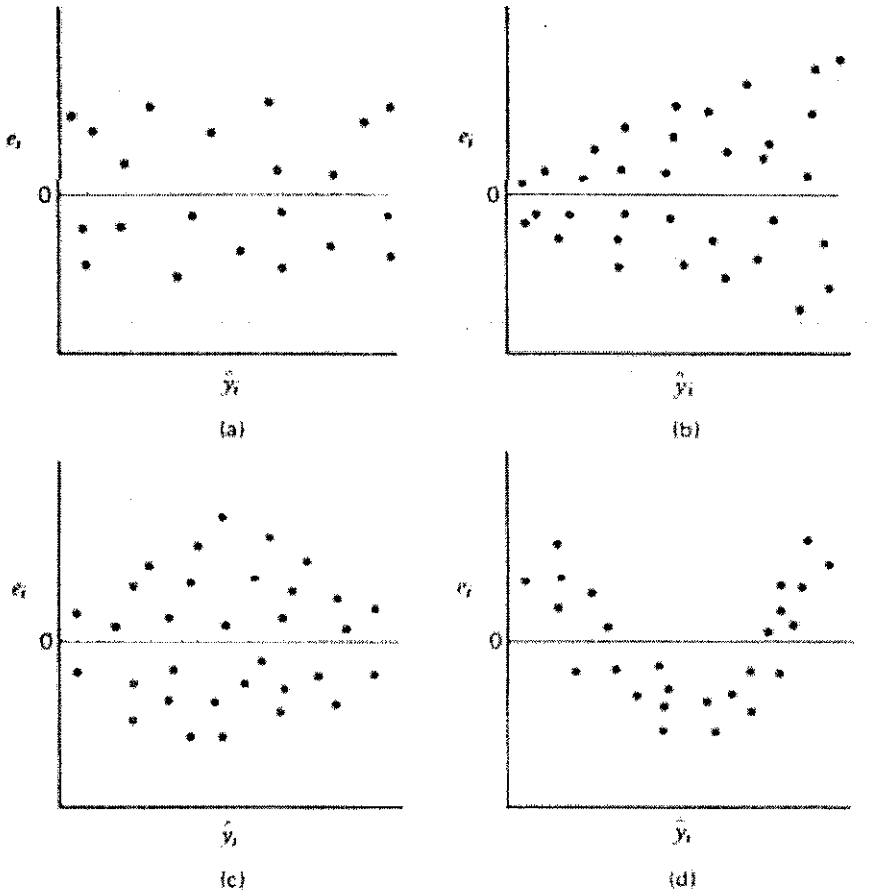
สร้างกราฟระหว่างค่าคาดหวังแบบปกติ ซึ่งอยู่บนแกน Y และค่า e_i ซึ่งอยู่บนแกน X ดังแสดงในรูปที่ 4.2 จะเห็นได้ว่าข้อมูลส่วนใหญ่ตกใกล้เคียงกับเส้นตรง มีเพียงค่าสังเกตตัวที่ 8 (มุมล่างซ้าย) ที่ตกห่างจากข้อมูลส่วนใหญ่ ซึ่งควรที่จะตรวจสอบต่อไปว่าค่าสังเกตดังกล่าวเป็นค่าที่ผิดปกติหรือไม่ และจากการตรวจสอบโดยใช้ Standardized residuals หรือ Studentized residuals ยังคงให้ผลในทำนองเดียวกัน



รูปที่ 4.2: Normal probability plot ของความคลาดเคลื่อน

4.2.2 กราฟของความคลาดเคลื่อน (e_i) กับค่าพยากรณ์ $(\hat{Y}_i)$

การสร้างกราฟระหว่างค่าความคลาดเคลื่อน e_i (อาจใช้ d_i หรือ r_i แทน e_i ได้) และค่าพยากรณ์มีประโยชน์ต่อการตรวจสอบข้อบกพร่องของโมเดลเช่นกัน พิจารณารูปที่ 4.3 (a) จะเห็นได้ว่าค่าความคลาดเคลื่อนมีการกระจายสม่ำเสมอเป็นแถบขนานกับแกนนอน (Horizontal band) ซึ่งเป็นลักษณะของความคลาดเคลื่อนที่สอดคล้องกับข้อกำหนดของการวิเคราะห์ ส่วนรูปที่ 4.3 (b)-(d) แสดงถึงข้อบกพร่องของโมเดลในลักษณะต่าง ๆ กัน โดยรูป ที่ 4.3 (b) และ (c) ชี้ว่าความแปรปรวนของความคลาดเคลื่อนไม่คงที่ และจะเห็นได้ว่าความคลาดเคลื่อนในรูป ที่ 4.3 (b) มีลักษณะคล้ายกรวย (Funnel) ที่บานออก แสดงว่าความแปรปรวนเป็นฟังก์ชันเพิ่มของค่าสังเกต Y นั่นคือ ความแปรปรวนเพิ่มขึ้น เมื่อค่าสังเกต Y เพิ่มขึ้น ในทางกลับกันหากความคลาดเคลื่อนมีลักษณะคล้ายกรวยที่บานออกแล้วจึงลู่เข้า แสดงว่าความแปรปรวนเป็นฟังก์ชันลดของค่าสังเกต Y นั่นคือ ความแปรปรวนเพิ่มขึ้น เมื่อค่าสังเกต Y ลดลง ส่วนรูปที่ 4.3 (c) เรียกว่า *Double-bow pattern* มักเกิดขึ้นเมื่อ Y เป็นสัดส่วนที่มีค่าอยู่ระหว่าง 0 และ 1 ซึ่งจะเห็นได้ว่าความแปรปรวนที่สัดส่วนมีค่าใกล้เคียง 0.5 มีค่ามากกว่าความแปรปรวนที่สัดส่วนมีค่าใกล้เคียง 0 หรือ 1 โดยปัญหาความแปรปรวนไม่คงที่ อาจแก้ไขได้โดยการแปลงค่า (Transformaiton) ของตัวแปรตาม หรือประมาณค่าพารามิเตอร์โดยใช้วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (Weighted least squares) สำหรับรูปที่ 4.3 (d) แสดงถึงลักษณะความสัมพันธ์ที่ไม่ใช่เชิงเส้นตรง (Nonlinearity) ทั้งนี้อาจแก้ไขได้โดยเพิ่มตัวแปรอิสระอื่นเข้า ไปในโมเดล เช่น เทอมกำลังสอง เป็นต้น หรือแปลงค่าของตัวแปรตามและตัวแปรอิสระตัวใดตัวหนึ่งหรือทั้งสองตัว

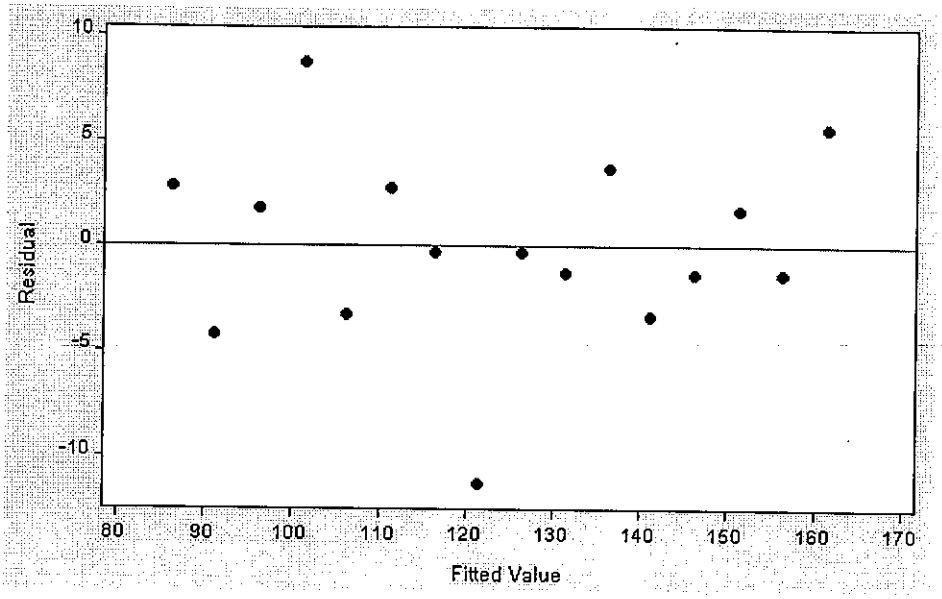


รูปที่ 4.3: ลักษณะของกราฟระหว่างค่าความคลาดเคลื่อน (e_i) กับค่าพยากรณ์ ($\hat{Y}_i$) (จาก Montgomery และ Peck (1992))

นอกจากนี้กราฟระหว่าง e_i และ $\hat{Y}_i$ อาจชี้ถึงค่าสังเกตที่มีความคลาดเคลื่อนสูงผิดปกติได้ ซึ่งควรที่จะมีการตรวจสอบต่อไปว่าค่าดังกล่าวเป็นข้อมูลที่ผิดปกติหรือไม่ ถ้าความคลาดเคลื่อนที่สูงผิดปกตินั้นเกิดขึ้นเมื่อ $\hat{Y}_i$ มีค่ามากหรือน้อยมาก ๆ (Extreme value) อาจเนื่องมาจากความแปรปรวนไม่คงที่หรือความสัมพันธ์ที่แท้จริงระหว่าง Y และ X ไม่ใช่ความสัมพันธ์เชิงเส้นตรงก็ได้

ตัวอย่างที่ 4.2 จากข้อมูลในตัวอย่างที่ 4.1 จงสร้างกราฟระหว่างความคลาดเคลื่อนที่เกิดจากสมการถดถอยเชิงเส้นตรงอย่างง่ายและค่าพยากรณ์

วิธีทำ จากตารางที่ 4.1 สร้างกราฟของความคลาดเคลื่อน (Residual plot) ได้ดังแสดงในรูปที่ 4.4 ซึ่งจะเห็นได้ว่ามีค่าสังเกตหนึ่งค่าที่มีความคลาดเคลื่อนต่ำผิดปกติ นอกจากค่าดังกล่าวแล้วพบว่าความคลาดเคลื่อนมีการกระจายอย่างสุ่มและไม่มีรูปแบบที่ชัดเจน



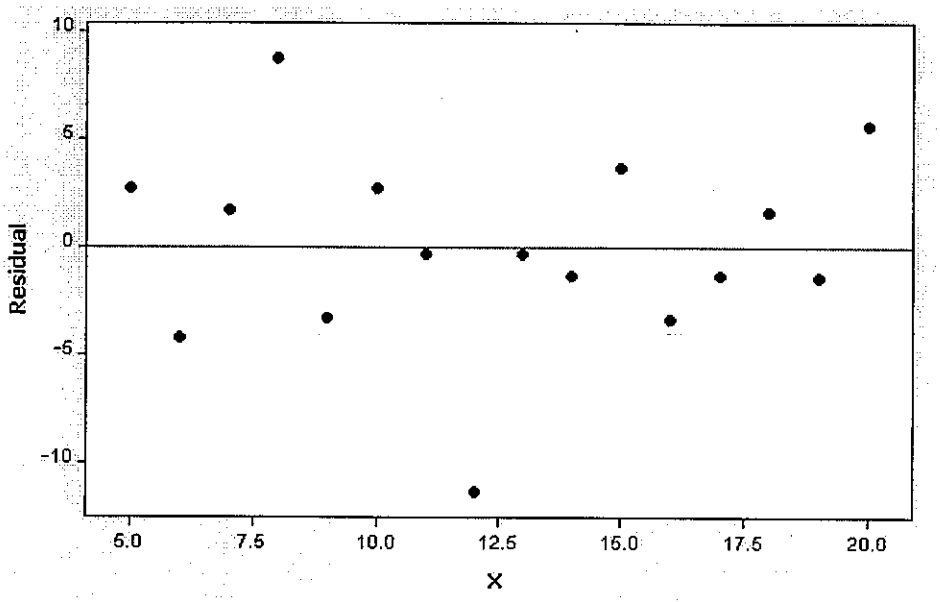
รูปที่ 4.4: กราฟระหว่างความคลาดเคลื่อนและค่าพยากรณ์ของข้อมูลยอดขายและค่าใช้จ่ายในการโฆษณา

4.2.3 กราฟของความคลาดเคลื่อน (e_i) กับตัวแปรอิสระแต่ละตัว (X_i)

กราฟระหว่างค่าความคลาดเคลื่อน e_i (d_i หรือ r_i) และตัวแปรอิสระแต่ละตัวสามารถใช้ตรวจสอบข้อกำหนดเบื้องต้นของการวิเคราะห์ได้เช่นเดียวกัน และมักแสดงลักษณะคล้ายคลึงกับรูปที่ 4.3 เพียงแต่แกนนอนเป็นค่าของตัวแปรอิสระ X แทนที่จะเป็นค่พยากรณ์ $\hat{Y}_i$ ทำนองเดียวกันถ้าความคลาดเคลื่อนมีการกระจายเป็นแถบขนานแกนนอน แสดงว่าความคลาดเคลื่อนมีลักษณะสอดคล้องกับข้อกำหนดเบื้องต้น แต่ถ้าความคลาดเคลื่อนมีรูปแบบคล้ายกรวยที่บานออกหรือลู่เข้า หรือ Double-bow pattern แสดงว่าความแปรปรวนไม่คงที่ และถ้ากราฟมีลักษณะเป็นแถบโค้ง ควรเพิ่มตัวแปรอิสระอื่นเข้าไปในโมเดลหรืออาจแปลงค่าของตัวแปรอิสระ

ตัวอย่างที่ 4.3 จากข้อมูลในตัวอย่างที่ 4.1 จงสร้างกราฟระหว่างความคลาดเคลื่อนที่เกิดจากสมการถดถอยเชิงเส้นตรงอย่างง่ายและค่าใช้จ่ายในการโฆษณา

วิธีทำ สร้างกราฟของความคลาดเคลื่อนกับค่าใช้จ่ายในการโฆษณา ได้ดังแสดงในรูปที่ 4.5 ซึ่งจะเห็นได้ว่า มีค่าสังเกตหนึ่งค่าที่มีความคลาดเคลื่อนต่ำผิดปกติ ซึ่งก็คือ ค่าสังเกตตัวที่ 8 นอกจากค่าดังกล่าวแล้วพบว่าความคลาดเคลื่อนมีลักษณะสอดคล้องกับข้อกำหนดเบื้องต้นของการวิเคราะห์

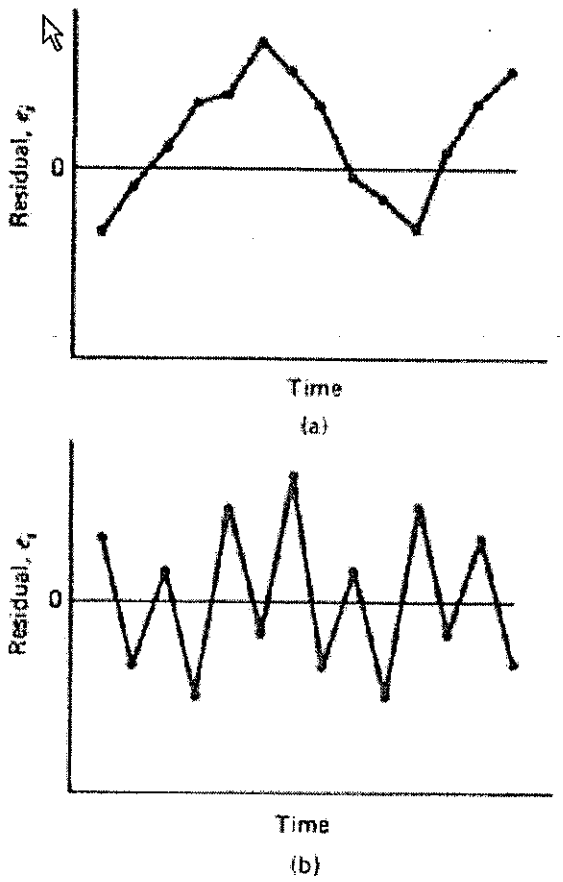


รูปที่ 4.5: กราฟระหว่างความคลาดเคลื่อนและค่าใช้จ่ายในการโฆษณา

4.2.4 กราฟของความคลาดเคลื่อน (e_i) กับลำดับเวลา (Time) หรือลำดับข้อมูล (Sequence)

ถ้าข้อมูลถูกเก็บตามลำดับเวลา การสร้างกราฟระหว่างความคลาดเคลื่อนกับลำดับเวลาจะช่วยในการตรวจสอบข้อกำหนดเบื้องต้นของการวิเคราะห์ได้เช่นกัน หากกราฟมีลักษณะคล้ายคลึงกับรูปที่ 4.3 (b)-(c) แสดงว่าความแปรปรวนมีค่าไม่คงที่ โดยเปลี่ยนแปลงตามลำดับเวลา และหากกราฟมีลักษณะคล้ายคลึงกับรูปที่ 4.3 (d) อาจแก้ไขโดยเพิ่มเทอมกำลังสองของเวลาเข้าไปในโมเดล นอกจากนี้กราฟประเภทนี้ยังช่วยในการตรวจสอบความคลาดเคลื่อนที่เวลาใดเวลาหนึ่งว่ามีความสัมพันธ์กับความคลาดเคลื่อนที่เวลาอื่นหรือไม่ ซึ่งความสัมพันธ์ระหว่างความคลาดเคลื่อนที่เวลาต่างกัน จะเรียกว่า *Autocorrelation*

จะเห็นได้ว่ารูปที่ 4.6 (a) และ (b) แสดงลักษณะของการเกิด Autocorrelation 2 ลักษณะ โดยรูปที่ 4.6 (a) ความคลาดเคลื่อนมีการเปลี่ยนแปลงต่อเนื่องกันตามระยะเวลาที่อยู่ติดกัน เรียกลักษณะความสัมพันธ์ระหว่างความคลาดเคลื่อนดังกล่าวนี้ว่า *Positive autocorrelation* แต่ถ้าความคลาดเคลื่อนมีการเปลี่ยนแปลงขึ้นลงสลับระหว่างค่าบวกและลบตามเวลาที่อยู่ติดกัน จะเรียกความสัมพันธ์นี้ว่า *Negative autocorrelation* ซึ่งการเกิด Autocorrelation จะแย้งกับข้อกำหนดเบื้องต้นของการวิเคราะห์เกี่ยวกับการเป็นอิสระกันของความคลาดเคลื่อน การตรวจสอบ Autocorrelation ยังสามารถทำได้โดยใช้สถิติทดสอบ ซึ่งจะกล่าวถึงในหัวข้อที่ 4.6 อย่างละเอียดต่อไป



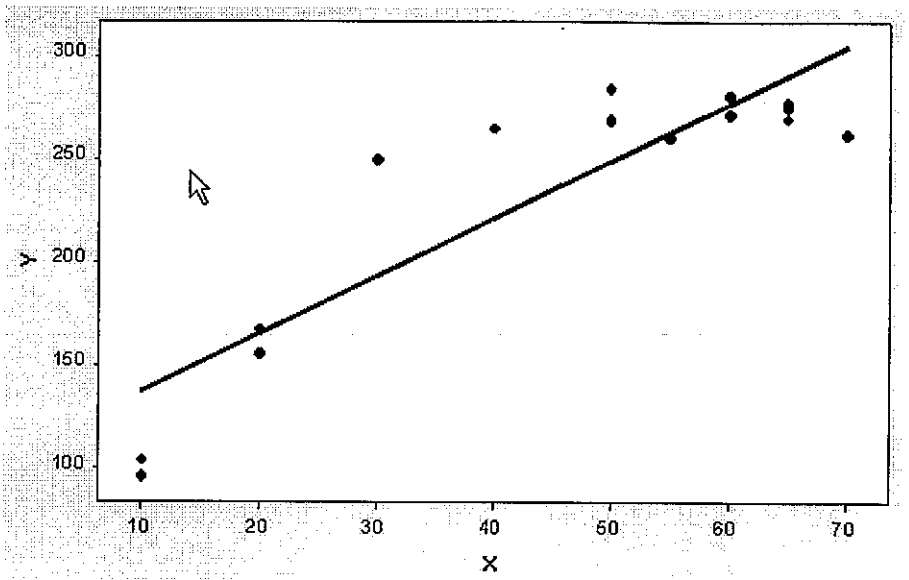
รูปที่ 4.6: กราฟระหว่างความคลาดเคลื่อนและเวลา (a) แสดง Positive autocorrelation (b) แสดง Negative autocorrelation (จาก Montgomery และ Peck (1992))

4.2.5 กราฟของความคลาดเคลื่อน (e_i) และตัวแปรอิสระอื่นที่ไม่อยู่ในโมเดล

กราฟระหว่างความคลาดเคลื่อนและตัวแปรอิสระที่ไม่อยู่ในโมเดลจะทำได้ก็ต่อเมื่อทราบค่าของตัวแปรอิสระที่ไม่อยู่ในโมเดลนั้น ๆ หากกราฟแสดงรูปแบบอย่างมีระบบ (Systematic pattern) แสดงว่าการเพิ่มตัวแปรอิสระตัวใหม่เข้าไปในโมเดล อาจทำให้โมเดลมีลักษณะที่เหมาะสมมากขึ้น

4.3 การทดสอบ Lack of Fit

การทดสอบ Lack of fit เป็นวิธีการทางสถิติที่ใช้ตรวจสอบความเหมาะสมของรูปแบบความสัมพันธ์ระหว่างตัวแปร ซึ่งการทดสอบดังกล่าวจะสมมติว่าข้อมูลมีการแจกแจงแบบปกติ เป็นอิสระกัน และมีความแปรปรวนคงที่ แต่ต้องการตรวจสอบว่าลักษณะความสัมพันธ์เชิงเส้นตรงหรือสมการกำลังหนึ่ง (First-order model) เหมาะสมกับข้อมูลหรือไม่ พิจารณารูปที่ 4.7 จะเห็นได้ว่าความสัมพันธ์ในข้อมูลมีลักษณะเป็นเส้นโค้งมากกว่าที่จะเป็นเส้นตรง ดังนั้นการสร้างสมการถดถอยเชิงเส้นตรงกับข้อมูลดังกล่าวจึงไม่เหมาะสม ในที่นี้จะกล่าวถึงการทดสอบ Lack of fit สำหรับโมเดลถดถอยเชิงเส้นตรงอย่างง่ายเท่านั้น ซึ่งสามารถขยายไปยังโมเดลถดถอย



รูปที่ 4.7: กราฟแสดง Lack of fit ของโมเดลถดถอยเชิงเส้นตรง

เชิงเส้นตรงแบบพหุได้เช่นเดียวกัน

การทดสอบ Lack of fit กำหนดว่าตัวแปรอิสระต้องมีค่าซ้ำกันอย่างน้อย 1 ระดับ โดยค่าสังเกตของ Y ที่เกิดซ้ำ ๆ กันในแต่ละระดับของ X เรียกว่า *Repeat observations* ซึ่ง Repeat observations ควรเกิดจากการทำซ้ำ (Replications) จริง ๆ ไม่ใช่เกิดจากการอ่านผลหรือวัดผลซ้ำ (Duplicate measurements)

สมมติในระดับที่ j ของตัวแปรอิสระ X ประกอบด้วยค่าของตัวแปรตาม n_j ค่า $(j = 1, 2, \ldots, m)$ ให้ Y_{ij} แทน ค่าสังเกตที่ i บนตัวแปรอิสระ X_j $(i = 1, 2, \ldots, n_j$ และ $j = 1, 2, \ldots, m)$ ดังนั้นจำนวนค่าสังเกตทั้งหมด คือ $n = \sum_{j=1}^{m} n_j$ สามารถแสดงได้ดังนี้

ระดับของตัวแปรอิสระ	ค่าสังเกต
X_1	$Y_{11}, Y_{21}, \ldots, Y_{n_1 1}$
X_2	$Y_{12}, Y_{22}, \ldots, Y_{n_2 2}$
$\vdots$	$\vdots$
X_m	$Y_{1m}, Y_{2m}, \ldots, Y_{n_m m}$

ตารางที่ 4.3: ลักษณะข้อมูลที่ใช้ในการทดสอบ Lack of fit

การทดสอบทำได้โดยแบ่งผลรวมกำลังสองของความคลาดเคลื่อนออกเป็น 2 ส่วน ดังนี้

$$SSE = SS_{PE} + SS_{LOF} \tag{4.9}$$

เมื่อ

SS_{PE} แทน ผลรวมกำลังสองเนื่องมาจากความคลาดเคลื่อนที่แท้จริง (Sum of squares due to pure error)

SS_{LOF} แทน ผลรวมกำลังสองเนื่องมาจากรูปแบบสมการไม่เหมาะสม (Sum of squares due to lack of fit)

พิจารณาค่าความคลาดเคลื่อนของค่าสังเกตตัวที่ i ของ X_j

$$Y_{ij} - \hat{Y}_j = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_j) \tag{4.10}$$

เมื่อ $\bar{Y}_j$ แทน ค่าเฉลี่ยของค่าสังเกตที่ระดับ X_j

ยกกำลังสองทั้งสองข้างแล้วหาผลรวม จะได้ว่า

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{m} n_j(\bar{Y}_j - \hat{Y}_j)^2$$

$$SSE \quad = \quad SS_{PE} \quad + \quad SS_{LOF}$$

เนื่องจากเทอม Cross-product มีค่าเท่ากับศูนย์ ถ้าข้อกำหนดเบื้องต้นเกี่ยวกับความแปรปรวนคงที่เป็นจริงแล้ว จะได้ว่า $SS_{PE} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$ เป็นค่าที่ใช้วัดความคลาดเคลื่อนที่แท้จริง โดยวัดความผันแปรของ Y ในแต่ละระดับของ X และไม่ขึ้นอยู่กับรูปแบบของโมเดล เพื่อให้สะดวกต่อการคำนวณค่า SS_{PE} ในแต่ละระดับของ X_j จะได้ว่า

$$\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2 \quad = \quad \sum_{i=1}^{n_j} Y_{ij}^2 - n_j\bar{Y}_j^2, \qquad j = 1, 2, \ldots, m \tag{4.11}$$

โดยในแต่ละระดับของ X_j มีจำนวนองศาความเป็นอิสระเท่ากับ $n_j - 1$ ดังนั้นจำนวนองศาความเป็นอิสระทั้งหมดที่สอดคล้องกับ SS_{PE} คือ

$$df_{PE} \quad = \quad \sum_{j=1}^{m}(n_j - 1) \quad = \quad n - m \tag{4.12}$$

เมื่อ df_{PE} แทน จำนวนองศาความเป็นอิสระของ Pure error

จะเห็นได้ว่า $SS_{LOF} = \sum_{j=1}^{m} n_j(\bar{Y}_j - \hat{Y}_j)^2$ เป็นค่าที่ได้จากการถ่วงน้ำหนักผลรวมของส่วนเบี่ยงเบนกำลังสองระหว่าง $\bar{Y}_j$ ในแต่ละระดับของ X กับค่าพยากรณ์ที่สอดคล้องกัน ถ้า $\hat{Y}_j$ มีค่าใกล้เคียงกับ $\bar{Y}_j$ แสดงว่าสมการถดถอยนั้นเหมาะสมแล้ว แต่ถ้า $\hat{Y}_j$ มีค่าเบี่ยงเบนไปจาก $\bar{Y}_j$ มาก ๆ อาจเป็นไปได้ที่สมการถดถอยนั้นไม่เหมาะสม เนื่องจากตัวแปรอิสระ X มีค่าที่แตกต่างกัน m ค่า ดังนั้นจำนวนองศาความเป็น

อิสระที่สอดคล้องกับ SS_{LOF} คือ $m-2$ โดยจำนวนองศาความเป็นอิสระที่สูญเสียไป 2 ค่านั้นเนื่องจากการประมาณค่า β_0 และ β_1 ใน $\hat{Y}_j$

สถิติที่ใช้ในการทดสอบ Lack of fit คือ

$$F_c = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \tag{4.13}$$

ทั้งนี้เนื่องจาก

$$E(MS_{PE}) = \sigma^2 \tag{4.14}$$

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{j=1}^{m} n_j \left[E(Y_j) - \beta_0 - \beta_1 X_j\right]^2}{m-2} \tag{4.15}$$

ถ้าสมการถดถอยที่แท้จริงเป็นสมการเชิงเส้นตรงแล้ว นั่นคือ $E(Y_j) = \beta_0 + \beta_1 X_j$ เทอมที่สองของสมการ (4.15) จะมีค่าเท่ากับ 0 และได้ว่า $E(MS_{LOF}) = \sigma^2$ แต่ถ้าสมการถดถอยที่แท้จริงไม่ใช่สมการเชิงเส้นตรงแล้ว นั่นคือ $E(Y_j) \neq \beta_0 + \beta_1 X_j$ และได้ว่า $E(MS_{LOF}) > \sigma^2$ นอกจากนี้ถ้าสมการถดถอยที่แท้จริงเป็นสมการเชิงเส้นตรงแล้ว จะได้ว่า

$$F_c \sim F_{m-2,\, n-m\,(\alpha)} \tag{4.16}$$

ดังนั้นถ้าสถิติทดสอบ $F_c > F_{m-2,\, n-m\,(\alpha)}$ สามารถสรุปได้ว่ารูปแบบของสมการถดถอยเชิงเส้นตรงนั้นไม่เหมาะสม ก็ควรที่จะค้นหารูปแบบสมการที่เหมาะสมกว่าต่อไป นอกจากนี้การทดสอบ Lack of fit ยังสามารถแสดงในตารางการวิเคราะห์ความแปรปรวนได้ดังนี้

กำหนดสมมติฐานของการทดสอบเป็น

H_0 : รูปแบบสมการถดถอยเหมาะสม

H_1 : รูปแบบสมการถดถอยไม่เหมาะสม

หรือ

$$H_0 : \quad E(Y) = \beta_0 + \beta_1 X$$

$$H_1 : \quad E(Y) \neq \beta_0 + \beta_1 X$$

ระดับนัยสำคัญของการทดสอบ 0.05

ANOVA

Source of variation	df	SS	MS	F
Regression	1	SSR	MSR	
Error	$n-2$	SSE	MSE	
Lack of fit	$m-2$	SS_{LOF}	MS_{LOF}	$F_c = \frac{MS_{LOF}}{MS_{PE}}$
Pure error	$n-m$	SS_{PE}	MS_{PE}	
Total	$n-1$	SST		

ค่าวิกฤติคือ $F_{m-2,n-m\,(\alpha)}$

โดยจะปฏิเสธ H_0 เมื่อ $F_c > F_{m-2,n-m\,(\alpha)}$

ตัวอย่างที่ 4.4 จากข้อมูลในตารางต่อไปนี้

X	10	10	20	20	30	40	50	50
Y	104	96	155	167	250	265	270	268
X	50	55	60	60	65	65	65	70
Y	284	260	272	281	275	278	270	262

ตารางที่ 4.4: ข้อมูลสำหรับตัวอย่างที่ 4.4

สร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายได้เป็น $\hat{Y} = 109.02 + 2.7953X$ ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่ารูปแบบโมเดลถดถอยเชิงเส้นตรงเหมาะสมหรือไม่

วิธีทำ สร้างแผนภาพการกระจายของข้อมูลชุดนี้ ดังแสดงในรูปที่ 4.7 ซึ่งจะเห็นได้ว่าความสัมพันธ์มีลักษณะเป็นเส้นโค้ง เพื่อที่จะทดสอบความเหมาะสมของรูปแบบความสัมพันธ์เชิงเส้นตรง กำหนดสมมติฐานของการทดสอบได้ดังนี้

H_0 : รูปแบบสมการถดถอยเชิงเส้นตรงอย่างง่ายเหมาะสม

H_1 : รูปแบบสมการถดถอยเชิงเส้นตรงอย่างง่ายไม่เหมาะสม

ระดับนัยสำคัญของการทดสอบ 0.05

คำนวณค่าต่าง ๆ ได้ดังนี้

$$SST = 62,858.44, \quad SSR = 50,008.14, \quad SSE = 12,850.30$$

จะเห็นได้ว่าตัวแปรอิสระ X มีค่าที่แตกต่างกัน 9 ระดับ แต่มีเพียง 5 ระดับที่เกิดซ้ำกัน คือ ที่ $X = 10,\ 20,\ 50,\ 60,$ และ 65 คำนวณค่า SS_{PE} ได้ดังนี้

ระดับของ X	ค่าสังเกต Y	$\bar{Y}_j$	$\sum_{i=1}^{n_j} Y_{ij}^2 - n_j\bar{Y}_j^2$	df
10	104, 96	100.00	32.00	1
20	155, 167	161.00	72.00	1
50	270, 268, 284	274.00	152.00	2
60	272, 281	276.50	40.50	1
65	275, 278, 270	274.33	32.67	2
		รวม	329.17	7

ตารางที่ 4.5: ตัวอย่างการคำนวณค่า SS_{PE}

สร้างตารางวิเคราะห์ความแปรปรวนได้ดังนี้

ANOVA

Source of variation	df	SS	MS	F
Regression	1	50,008.14	50,008.14	
Error	14	12,850.30	917.88	
Lack of fit	7	12,521.13	1,788.73	$F_c = 38.04$
Pure error	7	329.17	47.02	
Total	15	62,858.44		

ค่าวิกฤติคือ $F_{7,7\,(0.05)} = 3.79$

โดย SS_{LOF} คำนวณได้จาก

$$SS_{LOF} = SSE - SS_{PE} = 12,850.30 - 329.17 = 12,521.13$$

และมีจำนวนองศาความเป็นอิสระเท่ากับ $m - 2 = 9 - 2 = 7$

เนื่องจาก $F_c = 38.04$ มีค่ามากกว่าค่าวิกฤติ จึงปฏิเสธ H_0 ดังนั้นรูปแบบโมเดลถดถอยเชิงเส้นตรงอย่างง่ายไม่เหมาะสม ที่ระดับนัยสำคัญ 0.05

ข้อสังเกต

1. การทดสอบสมมติฐาน $H_0 : \beta_1 = 0$ (ไม่มีความสัมพันธ์เชิงเส้นตรงระหว่าง X และ Y) จะสมเหตุสมผลก็ต่อเมื่อรูปแบบโมเดลถดถอยเชิงเส้นตรงนั้นเหมาะสม ดังนั้นในการทดสอบสมมติฐานหรือประมาณช่วงความเชื่อมั่นจึงควรที่จะตรวจสอบความเหมาะสมของรูปแบบโมเดลก่อนเสมอ

2. เมื่อเกิด Lack of fit แสดงว่ารูปแบบโมเดลที่กำลังพิจารณานั้นไม่เหมาะสม ซึ่งอาจแก้ไขได้โดยค้นหารูปแบบอื่นของสมการที่เหมาะสมมากกว่า หรือแปลงข้อมูลเพื่อให้มีความสัมพันธ์เชิงเส้นตรง

3. หากไม่มีระดับของ X ที่มีค่าซ้ำกันเลย การทดสอบ Lack of fit โดยประมาณยังคงสามารถทำได้ โดยจัดกลุ่มข้อมูลที่มีระดับของ X ติดกันและมีค่าสังเกตใกล้เคียงกันไว้ด้วยกัน และเรียกกลุ่มข้อมูลที่ถูกจัดไว้ด้วยกันว่า *Pseudo-replicate*

4.4 การตรวจสอบการเท่ากันของความแปรปรวน (Tests for Constancy of Error Variance)

ดังได้กล่าวแล้วว่าการตรวจสอบความแปรปรวนของความคลาดเคลื่อนว่ามีค่าคงที่หรือไม่ สามารถทำได้โดยสร้างกราฟระหว่างความคลาดเคลื่อน e_i (d_i หรือ r_i) กับค่าพยากรณ์ $\hat{Y}$ หรือตัวแปรอิสระแต่ละตัว X_i ซึ่งจะเห็นได้ว่าถ้าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ เรียกว่า *Homoscedasticity* ความคลาดเคลื่อนควรที่จะมีการกระจายสม่ำเสมออย่างสุ่มรอบจุด 0 ดังแสดงในรูปที่ 4.3 (a) แต่ถ้าความแปรปรวนของความคลาดเคลื่อนมีค่าไม่คงที่ เรียกว่า *Heteroscedasticity* ความคลาดเคลื่อนจะมีการกระจายในลักษณะที่มีรูปแบบ (Pattern) ไม่สม่ำเสมอ ดังแสดงใน รูปที่ 4.3 (c)-(d) นอกจากนี้การทดสอบการเท่ากันของความแปรปรวนยังสามารถทำได้โดยใช้สถิติทดสอบ (?) ซึ่งในที่นี้จะนำเสนอ 2 วิธี ดังนี้

4.4.1 สถิติทดสอบของ Levene ที่มีการปรับค่า (Modified Levene Test)

เป็นสถิติที่พัฒนามาจากสถิติทดสอบของ Levene และไม่ขึ้นอยู่กับการแจกแจงแบบปกติของความคลาดเคลื่อน สามารถใช้ตรวจสอบความแปรปรวนของความคลาดเคลื่อนว่ามีลักษณะเพิ่มขึ้นหรือลดลงตามค่าของตัวแปรอิสระ X หรือไม่ โดยคำนึงถึงการกระจายของความคลาดเคลื่อนที่ได้จากตัวอย่าง การทดสอบนี้ทำได้โดยแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มข้อมูลที่มีค่า X อยู่ในระดับต่ำและระดับสูงตามลำดับ ถ้าความแปรปรวนของความคลาดเคลื่อนเพิ่มขึ้นหรือลดลงตามค่า X แล้ว ความแปรปรวนของความคลาดเคลื่อนในกลุ่มหนึ่งมีแนวโน้มสูงกว่าความแปรปรวนของความคลาดเคลื่อนในอีกกลุ่ม การคำนวณจะพิจารณาจากค่าเฉลี่ยของส่วนเบี่ยงเบนสัมบูรณ์ของความคลาดเคลื่อน (Mean absolute deviation of residuals) รอบมัธยฐานในแต่ละกลุ่มว่าแตกต่างกันอย่างมีนัยสำคัญหรือไม่

กำหนดให้

e_{i1} และ e_{i2} แทน ความคลาดเคลื่อนที่ i ที่ได้จากข้อมูลกลุ่มที่ 1 และ 2 ตามลำดับ

n_1 และ n_2 แทน ขนาดตัวอย่างในกลุ่มที่ 1 และ 2 ตามลำดับ โดย $n = n_1 + n_2$

$\tilde{e}_1$ และ $\tilde{e}_2$ แทน ค่ามัธยฐานของของความคลาดเคลื่อนในกลุ่มที่ 1 และ 2 ตามลำดับ

โดย

$$d_{i1} = \mid e_{i1} - \tilde{e}_1 \mid \quad \text{และ} \quad d_{i2} = \mid e_{i2} - \tilde{e}_2 \mid \tag{4.17}$$

เมื่อ d_{i1} และ d_{i2} แทน ส่วนเบี่ยงเบนสัมบูรณ์ของความคลาดเคลื่อนรอบมัธยฐานในกลุ่มที่ 1 และ 2 ตามลำดับ และให้ $\bar{d}_1$ และ $\bar{d}_2$ แทน ค่าเฉลี่ยตัวอย่างของ d_{i1} และ d_{i2} ตามลำดับ

สถิติทดสอบ

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{4.18}$$

โดย $S_p^2 = \dfrac{\sum_{i=1}^{n_1}(d_{i1}-\bar{d}_1)+\sum_{i=1}^{n_2}(d_{i2}-\bar{d}_2)}{n-2}$

เมื่อ t_L^* แทน สถิติทดสอบของ Levene ที่มีการปรับค่า

แม้ว่าส่วนเบี่ยงเบนสัมบูรณ์ของความคลาดเคลื่อนรอบมัธยฐานมักไม่มีการแจกแจงแบบปกติ แต่ถ้าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่และขนาดตัวอย่างทั้งสองกลุ่มไม่เล็กจนเกินไป พบว่า t_L^* มีการแจกแจงแบบทีโดยประมาณ ด้วยจำนวนองศาความเป็นอิสระ $n-2$ หาก $|t_L^*|$ ที่คำนวณได้มีค่ามาก แสดงว่าความแปรปรวนของความคลาดเคลื่อนมีค่าไม่คงที่

ตัวอย่างที่ 4.5 พิจารณาข้อมูลระหว่างค่าใช้จ่ายในการโฆษณา X (หน่วย: หมื่นบาท) และยอดขาย Y (หน่วย: หมื่นบาท) ของบริษัท 40 แห่ง ต่อไปนี้

บริษัทที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X	5.0	5.5	5.8	6.0	6.5	6.7	7.0	8.0	8.7	9.0	9.2	10.0	10.8	11.0
Y	89	90	85	87	93	93	98	110	108	103	108	114	112	116
บริษัทที่	15	16	17	18	19	20	21	22	23	24	25	26	27	28
X	11.4	11.6	12.0	12.5	13.0	13.1	13.6	14.0	15.0	15.7	16.0	16.5	17.0	17.6
Y	113	115	110	111	126	120	122	130	140	137	138	140	145	147
บริษัทที่	29	30	31	32	33	34	35	36	37	38	39	40		
X	18.0	18.2	18.9	19.0	19.1	19.9	20.0	20.2	20.5	21.0	21.6	22.0		
Y	153	155	158	155	160	167	169	166	163	175	171	177		

ตารางที่ 4.6: ข้อมูลยอดขายและค่าใช้จ่ายในการโฆษณา

สร้างสมการถดถอยได้เป็น $\hat{Y} = 57.152 + 5.2743X$ ที่ระดับนัยสำคัญ 0.05 จงตรวจสอบว่าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่หรือไม่

วิธีทำ แบ่งค่าสังเกต 40 ค่าออกเป็น 2 กลุ่ม โดยให้พิสัยของ X ทั้ง 2 กลุ่มมีค่าใกล้เคียงกัน ดังนี้ กลุ่มแรกประกอบด้วยค่าสังเกต 21 ค่าที่มีค่าใช้จ่ายในการโฆษณาตั้งแต่ 5 ถึง 13.6 หมื่นบาท และกลุ่มที่สองมีค่าสังเกต 19 ค่าที่มีค่าใช้จ่ายในการโฆษณาตั้งแต่ 14 ถึง 22 หมื่นบาท

กลุ่มที่ 1				
X	Y	e_{i1}	d_{i1}	$(d_{i1}-\bar{d}_1)^2$
5.0	89	5.4764	5.1944	37.1908
5.5	90	3.8393	3.5573	19.9032
5.8	85	-2.7430	-3.0250	4.4986
6.0	87	-1.7978	-2.0798	1.3826
6.5	93	1.5650	1.2830	4.7831
6.7	93	0.5102	0.2282	1.2819
7.0	98	3.9279	3.6459	20.7017
8.0	110	10.6537	10.3717	127.1404
8.7	108	4.9617	4.6797	31.1775
9.0	103	-1.6206	-1.9026	0.9972
9.2	108	2.3246	2.0426	8.6822
10.0	114	4.1051	3.8231	22.3459
10.8	112	-2.1143	-2.3963	2.2268
11.0	116	0.8309	0.5489	2.1109
11.4	113	-4.2788	-4.5608	13.3723
11.6	115	-3.3337	-3.6157	7.3531
12.0	110	-10.4434	-10.7254	96.4592
12.5	111	-12.0805	-12.3625	131.2971
13.0	126	0.2824	0.0004	0.8179
13.1	120	-6.2450	-6.5270	31.6186
13.6	122	-6.8822	-7.1642	39.1898
		รวม	-18.9841	604.5310

ตารางที่ 4.7: การคำนวณส่วนเบี่ยงเบนมาตรฐานสัมบูรณ์ของข้อมูลในกลุ่ม 1 โดยที่ $\bar{e}_1 = 0.282$ และ $\bar{d}_1 = -0.904$

จากตารางที่ 4.7 และ 4.8 คำนวณค่าความแปรปรวนร่วมได้ดังนี้

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^{n_1}(d_{i1}-\bar{d}_1)+\sum_{i=1}^{n_2}(d_{i2}-\bar{d}_2)}{n-2} \\ &= \frac{604.5310+212.2080}{40-2} \\ &= 21.4931 \end{aligned}$$

และได้ค่าสถิติทดสอบเป็น

$$\begin{aligned} t_L^* &= \frac{\bar{d}_1-\bar{d}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \\ &= \frac{-0.904+0.224}{\sqrt{21.4931\left(\frac{1}{21}+\frac{1}{19}\right)}} \\ &= -0.4633 \end{aligned}$$

กลุ่มที่ 2

X	Y	e_{i2}	d_{i2}	$(d_{i2}-\bar{d}_2)^2$
14.0	130	-0.99188	-1.90288	2.8186
15.0	140	3.73387	2.82287	9.2834
15.7	137	-2.95811	-3.86911	13.2868
16.0	138	-3.54039	-4.45139	17.8708
16.5	140	-4.17751	-5.08851	23.6635
17.0	145	-1.81464	-2.72564	6.2582
17.6	147	-2.97919	-3.89019	13.4410
18.0	153	0.91110	0.00010	0.0502
18.2	155	1.85625	0.94525	1.3672
18.9	158	1.16427	0.25327	0.2278
19.0	155	-2.36315	-3.27415	9.3034
19.1	160	2.10942	1.19842	2.0233
19.9	167	4.89002	3.97902	17.6654
20.0	169	6.36259	5.45159	32.2123
20.2	166	2.30774	1.39674	2.6268
20.5	163	-2.27454	-3.18554	8.7707
21.0	175	7.08834	6.17734	40.9771
21.6	171	-0.07622	-0.98722	0.5825
22.0	177	3.81408	2.90308	9.7786
		รวม	-4.2469	212.2080

ตารางที่ 4.8: การคำนวณส่วนเบี่ยงเบนมาตรฐานสัมบูรณ์ของข้อมูลในกลุ่ม 2 โดยที่ $\tilde{e}_2 = 0.911$ และ $\bar{d}_2 = -0.224$

ค่าวิกฤติคือ $t_{38\,(0.025)} = 2.024$

เนื่องจาก $\mid t_L^* \mid = 0.4633$ มีค่าน้อยกว่าค่าวิกฤติ ดังนั้นสามารถสรุปได้ว่าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่และไม่เปลี่ยนแปลงตามค่า X ที่ระดับนัยสำคัญ 0.05

หากความแปรปรวนของความคลาดเคลื่อนเปลี่ยนแปลงตามค่า X ในลักษณะที่ซับซ้อนมากกว่าที่จะเพิ่มขึ้นหรือลดลงเพียงอย่างเดียว การทดสอบการเท่ากันของความแปรปรวนอาจทำได้โดยแบ่งกลุ่มข้อมูลมากกว่าสองกลุ่ม และอาศัยหลักของวิธีวิเคราะห์ความแปรปรวนเพื่อทดสอบสมมติฐานดังกล่าว โดยใช้สถิติของ Levene นอกจากนี้สถิติที่ใช้ทดสอบการเท่ากันของความแปรปรวนที่มีความแกร่ง มักเป็นที่พึงปรารถนาของนักสถิติและนักวิเคราะห์ เนื่องจากปัญหาของการแจกแจงที่ไม่ใช่แบบปกติและความแปรปรวนไม่คงที่มักเกิดควบคู่กัน

4.4.2 สถิติทดสอบของ Breusch-Pagan (Breusch-Pagan Test)

การทดสอบของ Breusch-Pagan จัดเป็นการทดสอบสำหรับตัวอย่างที่มีขนาดใหญ่ (Large-sample test) สมมติให้ความคลาดเคลื่อนเป็นอิสระกันและมีการแจกแจงแบบปกติ หากความแปรปรวนของความคลาดเคลื่อน (ϵ_i)

ซึ่งแทนด้วย σ_i^2 มีความสัมพันธ์กับระดับของ X ตามรูปแบบต่อไปนี้

$$\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_i \tag{4.19}$$

จะเห็นได้ว่า σ_i^2 เป็นฟังก์ชันเพิ่มหรือฟังก์ชันลดของ X ขึ้นอยู่กับเครื่องหมายของ γ_1 หาก $\gamma_1 = 0$ แสดงว่าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ ดังนั้นการทดสอบการเท่ากันของความแปรปรวนด้วยวิธีของ Breusch-Pagan จะกำหนดสมมติฐานของการทดสอบดังนี้

$$H_0 : \gamma_1 = 0 \qquad vs. \qquad H_1 : \gamma_1 \neq 0$$

จากนั้นสร้างสมการถดถอยระหว่างค่าความคลาดเคลื่อนกำลังสอง (e_i^2) และ X_i แล้วคำนวณค่า SSR^* ได้สถิติทดสอบของ Breusch-Pagan ดังนี้

$$X_{BP}^2 = \frac{SSR^*/2}{(SSE/n)^2} \tag{4.20}$$

เมื่อ

SSR^* แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอยระหว่าง e_i^2 และ X_i

SSE แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอยระหว่าง Y_i และX_i

ถ้า $H_0 : \gamma_1 = 0$ เป็นจริงและตัวอย่างมีขนาดใหญ่พอ จะได้ว่า X_{BP}^2 มีการแจกแจงแบบไคสแควร์ ด้วยจำนวนองศาความเป็นอิสระเท่ากับ 1 ดังนั้นจะปฏิเสธ H_0 ถ้า X_{BP}^2 มีค่ามากกว่าค่าวิกฤติ

ตัวอย่างที่ 4.6 ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าความแปรปรวนของข้อมูลในตัวอย่างที่ 4.5 มีค่าคงที่หรือไม่ โดยใช้วิธีของ Breusch-Pagan

วิธีทำ จากตัวอย่างที่ 4.5 ได้สมการถดถอยเป็น $\hat{Y} = 57.152 + 5.2743X$ และคำนวณค่าต่าง ๆ ได้ดังนี้

$$SST = 30,700.975, \quad SSR = 29,867.132, \quad SSE = 833.843$$

จากตารางที่ 4.9 สร้างสมการถดถอยระหว่าง e_i^2 และ X_i และคำนวณค่าต่าง ๆ ได้ดังนี้

$$\hat{\gamma}_0 = 29.734, \qquad \hat{\gamma}_1 = -0.650,$$

$$SST^* = 41,181.038, \qquad SSR^* = 454.178, \qquad SSE^* = 40,726.860$$

X	Y	e_i	e_i^2	X	Y	e_i	e_i^2
5.0	89	5.4764	29.9912	13.6	122	-6.8822	47.3643
5.5	90	3.8393	14.7402	14.0	130	-0.9919	0.9838
5.8	85	-2.7430	7.5239	15.0	140	3.7339	13.9418
6.0	87	-1.7978	3.2322	15.7	137	-2.9581	8.7504
6.5	93	1.5650	2.4493	16.0	138	-3.5404	12.5343
6.7	93	0.5102	0.2603	16.5	140	-4.1775	17.4516
7.0	98	3.9279	15.4285	17.0	145	-1.8146	3.2929
8.0	110	10.6537	113.5004	17.6	147	-2.9792	8.8756
8.7	108	4.9617	24.6182	18.0	153	0.9111	0.8301
9.0	103	-1.6206	2.6263	18.2	155	1.8563	3.4457
9.2	108	2.3246	5.4035	18.9	158	1.1643	1.3555
10.0	114	4.1051	16.8522	19.0	155	-2.3632	5.5845
10.8	112	-2.1143	4.4701	19.1	160	2.1094	4.4497
11.0	116	0.8309	0.6904	19.9	167	4.8900	23.9123
11.4	113	-4.2788	18.3082	20.0	169	6.3626	40.4826
11.6	115	-3.3337	11.1133	20.2	166	2.3077	5.3257
12.0	110	-10.4434	109.0639	20.5	163	-2.2745	5.1735
12.5	111	-12.0805	145.9383	21.0	175	7.0883	50.2445
13.0	126	0.2824	0.0797	21.6	171	-0.0762	0.0058
13.1	120	-6.2450	39.0006	22.0	177	3.8141	14.5472

ตารางที่ 4.9: แสดงค่า e_i และ e_i^2 ของสมการถดถอยระหว่าง Y_i และ X_i

กำหนดสมมติฐานของการทดสอบ

$$H_0 : \gamma_1 = 0 \qquad vs. \qquad H_1 : \gamma_1 \neq 0$$

คำนวณสถิติทดสอบ

$$\begin{aligned} X^2_{BP} &= \frac{SSR^*/2}{(SSE/n)^2} \\ &= \frac{454.178/2}{(833.843/40)^2} \\ &= 0.5226 \end{aligned}$$

ค่าวิกฤติ $\chi^2_{1\,(0.05)} = 3.84$

เนื่องจากค่า $X^2_{BP} < 3.84$ ดังนั้นไม่มีเหตุผลเพียงพอที่จะปฏิเสธ H_0 นั่นคือ ความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ ที่ระดับนัยสำคัญ 0.05

หมายเหตุ การทดสอบความแปรปรวนว่ามีค่าคงที่หรือไม่ด้วยวิธีของ Breusch-Pagan สามารถใช้เมื่อรูปแบบความสัมพันธ์ระหว่าง e_i และ X_i มีลักษณะอื่นแตกต่างจากที่นำเสนอได้เช่นกัน

4.4.3 การแก้ไขเมื่อความแปรปรวนของความคลาดเคลื่อนมีค่าไม่คงที่

เป็นที่ทราบแล้วว่าความแปรปรวนคงที่เป็นข้อกำหนดพื้นฐานของการวิเคราะห์การถดถอย ดังนั้นการตรวจสอบและแก้ไขปัญหาความแปรปรวนไม่คงที่จึงเป็นสิ่งสำคัญ หากปัญหาดังกล่าวไม่ได้รับการแก้ไขแล้ว ถึงแม้ว่าตัวประมาณกำลังสองยังคงเป็นตัวประมาณที่ไม่เอนเอียง แต่จะขาดคุณสมบัติเกี่ยวกับความแปรปรวนต่ำสุด และส่งผลให้ค่าสัมประสิทธ์ถดถอยมีความคลาดเคลื่อนมาตรฐานสูงกว่าที่ควรจะเป็น ซึ่งการแปลงข้อมูลมักทำให้ได้ค่าประมาณพารามิเตอร์ที่แม่นยำและมีความไว (Sensitivity) ต่อการทดสอบทางสถิติมากขึ้น วิธีการแก้ไขเมื่อความแปรปรวนของความคลาดเคลื่อนมีค่าไม่คงที่ทำได้หลายวิธี ดังนี้

1. แปลงค่าของตัวแปรตาม Y

ความสัมพันธ์ระหว่าง σ^2 และ $E(Y)$	วิธีแปลงข้อมูล
$\sigma^2 \propto$ ค่าคงที่	$Y' = Y$
$\sigma^2 \propto E(Y)$	$Y' = \sqrt{Y}$
$\sigma^2 \propto E(Y)[1 - E(Y)]$	$Y' = \sin^{-1}(\sqrt{Y})$
$\sigma^2 \propto [E(Y)]^2$	$Y' = \ln Y$
$\sigma^2 \propto [E(Y)]^3$	$Y' = Y^{-1/2}$
$\sigma^2 \propto [E(Y)]^4$	$Y' = Y^{-1}$

ตารางที่ 4.10: วิธีแปลงข้อมูลเพื่อให้ความแปรปรวนคงที่ (Variance-stabilizing transformation)

โดยทั่วไปแล้วความแปรปรวนของข้อมูลไม่คงที่มักเกิดขึ้นในกรณีที่ Y มีการแจกแจงความน่าจะเป็นที่ความแปรปรวนเป็นฟังก์ชันของค่าเฉลี่ย เช่น ถ้า Y เป็นตัวแปรสุ่มแบบปัวซองส์ (Poisson random variable) ความแปรปรวนของ Y มีค่าเท่ากับค่าเฉลี่ยของ Y เนื่องจากสมการถดถอยจะแสดงความสัมพันธ์ระหว่างค่าเฉลี่ยของ Y กับ X ดังนั้นความแปรปรวนของ Y จะแปรผันตาม X ไปด้วย ซึ่งสามารถแก้ไขได้โดยสร้างสมการถดถอยระหว่าง $Y' = \sqrt{Y}$ กับ X เนื่องจากความแปรปรวนของ $\sqrt{Y}$ เป็นอิสระจากค่าเฉลี่ยของ Y หรือถ้า Y เป็นสัดส่วน $(0 \leq Y \leq 1)$ และกราฟของความคลาดเคลื่อนกับ $\hat{Y}$ มีลักษณะ Double-bow พบว่าการแปลงข้อมูลโดยใช้ arcsin เป็นวิธีที่เหมาะสม ซึ่งวิธีแปลงข้อมูลเพื่อทำให้ความแปรปรวนคงที่ที่นิยมใช้มีหลายวิธี สามารถสรุปได้ดังตารางที่ 4.10 นอกจากนี้ยังอาจใช้ความรู้ทางด้านทฤษฎีประกอบกับประสบการณ์เดิมช่วยในการเลือกวิธีแปลงข้อมูลที่เหมาะสมได้อีกด้วย

เมื่อแปลงค่าของตัวแปรตาม Y แล้ว จะทำให้ค่าพยากรณ์มีสเกลเปลี่ยนตามไปด้วย และบ่อยครั้งจำเป็นที่จะต้องเปลี่ยนค่าพยากรณ์ที่ได้กลับไปอยู่ในหน่วยเดิม แต่การแปลงกลับดังกล่าวจะทำให้ค่าพยากรณ์ที่ได้เป็นค่าประมาณของมัธยฐานของการแจกแจงของตัวแปรตามแทนที่จะเป็นค่าประมาณของค่าเฉลี่ย แต่อย่างไรก็ตามช่วงแห่งความเชื่อมั่นหรือช่วงแห่งการพยากรณ์สามารถแปลงกลับได้โดยตรง เนื่องจากการประมาณช่วงดังกล่าว

เป็นค่าเปอร์เซนไทล์ของการแจกแจง และเปอร์เซนไทล์ไม่ถูกกระทบด้วยวิธีการแปลงข้อมูล

หมายเหตุ บางครั้งการแปลงค่าของ Y อาจต้องบวกค่าคงที่เข้าไปด้วย เช่น เมื่อ Y มีค่าติดลบ การแปลงโดยใช้ลอการิทึม (Logarithm transformation) จะเปลี่ยนจุดเริ่มต้นของ Y และทำให้ค่าสังเกตมีค่ามากกว่าศูนย์ทุกค่า นั่นคือ $Y' = \log_{10}(Y+k)$ หรือ $\ln(Y+k)$ เมื่อ k แทน ค่าคงที่ที่เหมาะสม

2. แปลงค่าของตัวแปรตามด้วยวิธีของ Box-Cox (Box-Cox Transformation)

การหาวิธีแปลงค่าของตัวแปรตามที่เหมาะสมโดยพิจารณาจากกราฟของความคลาดเคลื่อนเพียงอย่างเดียวอาจทำได้ยาก วิธีของ Box-Cox เป็นวิธีหนึ่งที่ใช้ค้นหากำลังของการแปลงข้อมูลที่เหมาะสม โดยประมาณจากกลุ่มของ *Power transformation* ซึ่งอยู่ในรูป

$$Y' = Y^{\lambda} \tag{4.21}$$

เมื่อ λ แทน พารามิเตอร์ที่ต้องการประมาณค่า เช่น

$$\begin{aligned}
\lambda &= 2, & Y' &= Y^2 \\
\lambda &= 0.5, & Y' &= \sqrt{Y} \\
\lambda &= 0, & Y' &= \ln Y \\
\lambda &= -0.5, & Y' &= \frac{1}{\sqrt{Y}} \\
\lambda &= -1, & Y' &= \frac{1}{Y} \\
\lambda &= -2, & Y' &= \frac{1}{Y^2}
\end{aligned}$$

โมเดลถดถอยเชิงเส้นตรงอย่างง่ายเมื่อตัวแปรตามถูกแปลงค่าและเป็นสมาชิกของ Power transformation สามารถเขียนได้ดังนี้

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.22}$$

จะเห็นได้ว่าโมเดล (4.22) มีพารามิเตอร์ λ เพิ่มขึ้นมา กระบวนการของ Box-Cox จะใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum likelihood method) ในการประมาณพารามิเตอร์ λ รวมทั้งพารามิเตอร์ β_0, β_1 และ σ^2 ดังนั้นฟังก์ชันภาวะน่าจะเป็นสำหรับโมเดลถดถอย (4.22) มีรูปแบบดังนี้

$$L(\beta_0, \beta_1, \sigma^2, \lambda) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i^{\lambda} - \beta_0 - \beta_1 X_i)^2\right\} \tag{4.23}$$

ตัวประมาณของพารามิเตอร์ β_0, β_1, σ^2 และ λ ที่ทำให้ฟังก์ชัน (4.23) มีค่าสูงสุด จะถูกเรียกว่า *ตัวประมาณ*

ภาวะน่าจะเป็นสูงสุด (Maximum likelihood estimator: MLE) วิธีของ Box-Cox จะค้นหา $\hat{\lambda}$ ซึ่งเป็น MLE เพื่อใช้ในการแปลงข้อมูลใน Power transformation ขั้นตอนการค้นหา $\hat{\lambda}$ จะใช้หลักการค้นหาเชิงตัวเลข (Numerical search) โดยสมมติค่าที่เป็นไปได้ของ λ ขึ้นมาชุดหนึ่ง เช่น $\lambda = -1, \lambda = -0.9, \ldots, \lambda = 1.0$ เป็นต้น และในแต่ละค่าของ λ จะแปลงค่าสังเกต Y_i^{λ} ให้อยู่ในรูปคะแนนมาตรฐาน เพื่อให้ขนาดของ SSE ไม่ขึ้นอยู่กับค่าของ λ ดังนี้

$$Y_i^{(\lambda)} = \begin{cases} K_1(Y_i^{\lambda} - 1) & , \lambda \neq 0 \\ K_2(\ln Y_i) & , \lambda = 0 \end{cases} \tag{4.24}$$

เมื่อ

$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}$ แทน ค่าเฉลี่ยเรขาคณิต (Geometric mean) ของค่าสังเกต Y_i

$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$

หลังจากที่ได้ค่าสังเกต $Y_i^{(\lambda)}$ สำหรับ λ ที่กำหนดแล้ว สร้างสมการถดถอยระหว่าง $Y_i^{(\lambda)}$ กับตัวแปรอิสระ ทำเช่นนี้ไปเรื่อย ๆ จนครบทุกค่าของ λ จากนั้นเลือก λ ที่ทำให้ SSE มีค่าต่ำสุด ซึ่งอาจพิจารณาได้จากกราฟระหว่าง SSE กับ λ โดยค่าของ λ ที่ทำให้ SSE มีค่าต่ำสุดก็คือตัวประมาณแบบภาวะน่าจะเป็นสูงสุดที่ต้องการนั่นเอง

โดยทั่วไปวิธีของ Box-Cox จะใช้เพื่อหาวิธีแปลงข้อมูลโดยประมาณ ดังนั้นอาจไม่จำเป็นที่จะต้องระบุค่าของ $\hat{\lambda}$ ที่ละเอียด แต่จะใช้ค่าที่ใกล้เคียงและง่ายต่อการแปลผลแทน เช่น ถ้าได้ค่า λ ใกล้เคียง 0 อาจเลือกที่จะไม่แปลงข้อมูลก็ได้ นอกจากนี้ตัวประมาณแบบภาวะน่าจะเป็นสูงสุดของ λ ที่ได้จากวิธีของ Box-Cox มีความผันแปรที่เกิดจากการสุ่มตัวอย่างเกิดขึ้น แต่ SSE มีค่าค่อนข้างคงที่ในบริเวณรอบ ๆ ค่าประมาณ ดังนั้นจึงสมเหตุสมผลที่จะใช้ค่าที่ใกล้เคียงของ λ ในการแปลงข้อมูล เพื่อให้ง่ายต่อการทำความเข้าใจ หลังจากได้ $\hat{\lambda}$ ที่เหมาะสมแล้ว ควรสร้างแผนภาพการกระจายและมีการวิเคราะห์ความคลาดเคลื่อนควบคู่ไปกับการสร้างสมการเพื่อตรวจสอบความเหมาะสมของวิธีแปลงข้อมูลที่ได้จากวิธีของ Box-Cox เสมอ เมื่อรับเอาโมเดลที่มีการแปลงค่าตัวแปรตามไปใช้ ตัวประมาณพารามิเตอร์ที่ได้จากวิธีกำลังสองน้อยที่สุด มีคุณสมบัติสอดคล้องกับค่าสังเกตที่มีการแปลงค่าแล้ว ไม่ใช่ค่าสังเกตเดิม

ตัวอย่างที่ 4.7 จากข้อมูลที่กำหนดให้ จงหาวิธีแปลงข้อมูลที่เหมาะสมเพื่อให้ความแปรปรวนคงที่ โดยใช้วิธีของ Box-Cox

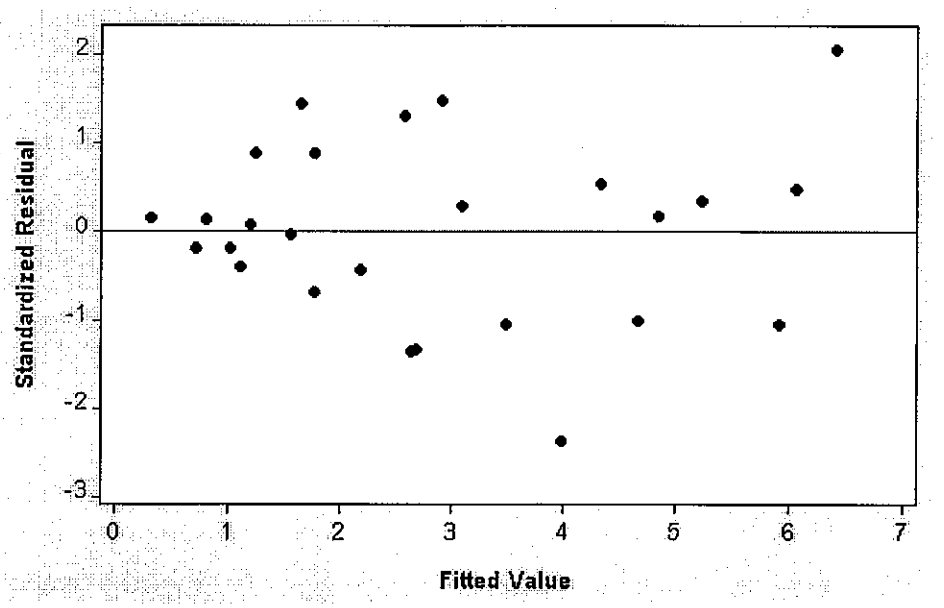
Y	1.50	0.56	0.48	0.71	2.62	3.56	4.65	9.42	5.26	6.77	5.76	5.13	3.17
X	667	280	1000	498	570	1144	985	2177	1085	2066	1806	1688	735
Y	4.35	3.08	0.42	1.02	1.80	0.69	1.31	0.48	1.48	5.20	0.56	3.92	0.23
X	2018	1631	402	434	1264	733	555	528	862	1531	1017	698	1422

ตารางที่ 4.11: ตัวอย่างของข้อมูลที่มีความแปรปรวนไม่คงที่

วิธีทำ สร้างสมการถดถอยของข้อมูลในตารางที่ 4.11 ได้เป็น

$$\hat{Y} = -0.589 + 0.0032X$$

และได้กราฟระหว่างความคลาดเคลื่อนกับค่าพยากรณ์ของสมการข้างต้นดังนี้

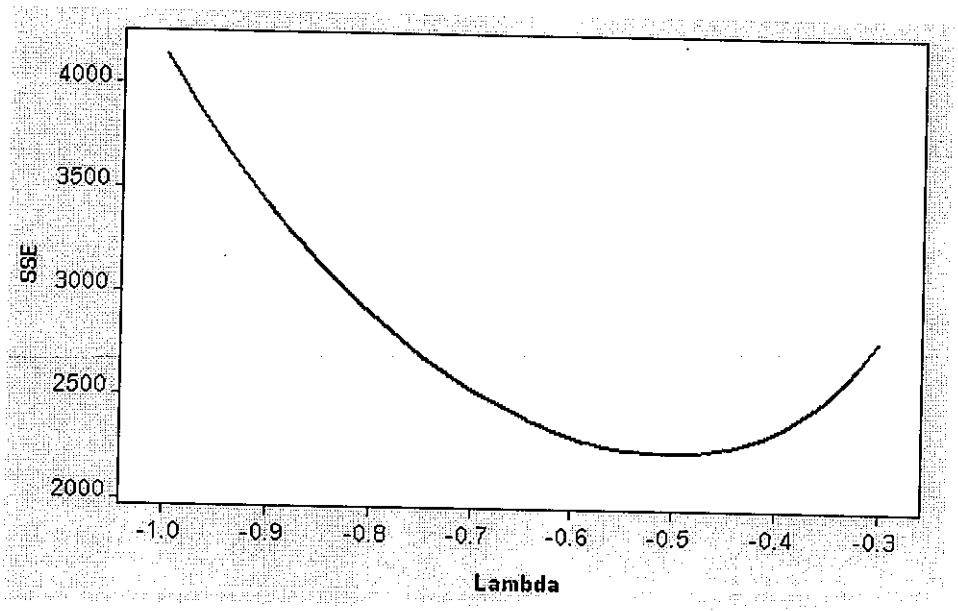


รูปที่ 4.8: กราฟระหว่างความคลาดเคลื่อนกับค่าพยากรณ์ของข้อมูลในตารางที่ 4.11

จากรูปที่ 4.8 จะเห็นได้ว่าปัญหาที่เกิดกับข้อมูลชุดนี้ก็คือ ความแปรปรวนของข้อมูลไม่คงที่ โดยเพิ่มตามค่าเฉลี่ย จากนั้นใช้วิธีของ Box-Cox เพื่อหากำลังของการแปลงข้อมูลที่เหมาะสม โดยกำหนดค่า $\lambda = -1.0, -0.95, -0.90, \ldots, -0.35, -0.30$ และคำนวณค่า SSE ที่สอดคล้องกับ λ แต่ละค่าได้ดังนี้

λ	-1.00	-0.95	-0.90	-0.85	-0.80	-0.75	-0.70	-0.65	-0.60
SSE	4142.70	3752.86	3419.59	3136.12	2896.93	2697.71	2535.26	2407.59	2314.05
λ	-0.55	-0.50	-0.45	-0.40	-0.35	-0.30			
SSE	2255.66	2235.70	2260.92	2343.77	2506.96	2793.59			

ตารางที่ 4.12: ค่า SSE ที่สอดคล้องกับ λ ที่กำหนด



รูปที่ 4.9: กราฟระหว่าง SSE และ λ

พิจารณารูปที่ 4.9 จะเห็นได้ว่าค่า λ ที่ทำให้ SSE มีค่าต่ำสุดก็คือ -0.5 นั่นคือตัวประมาณภาวะน่าจะเป็นสูงสุดมีค่าเป็น $\hat{\lambda} = -0.5$ ดังนั้นวิธีของ Box-Cox ชี้ว่ากำลังของการแปลงข้อมูลคือ -0.5

3. การประมาณพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (Weighted Least Square Estimation)

การประมาณค่าสัมประสิทธิ์ถดถอยสำหรับโมเดลที่มีความแปรปรวนไม่คงที่สามารถทำได้โดยใช้วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก โดยคูณส่วนเบี่ยงเบนระหว่างค่าสังเกตและค่าคาดหวังของตัวแปรตาม Y ด้วยน้ำหนัก (Weight) w_i และได้ผลรวมกำลังสองแบบถ่วงน้ำหนักสำหรับโมเดลถดถอยเชิงเส้นตรงอย่างง่าย ดังนี้

$$SSE = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_i)^2 \tag{4.25}$$

จากนั้นหาอนุพันธ์เทียบกับ β_0 และ β_1 แล้วให้เท่ากับศูนย์ จะได้สมการปกติเป็น

$$b_0 \sum_{i=1}^{n} w_i + b_1 \sum_{i=1}^{n} w_i X_i = \sum_{i=1}^{n} w_i Y_i \tag{4.26}$$

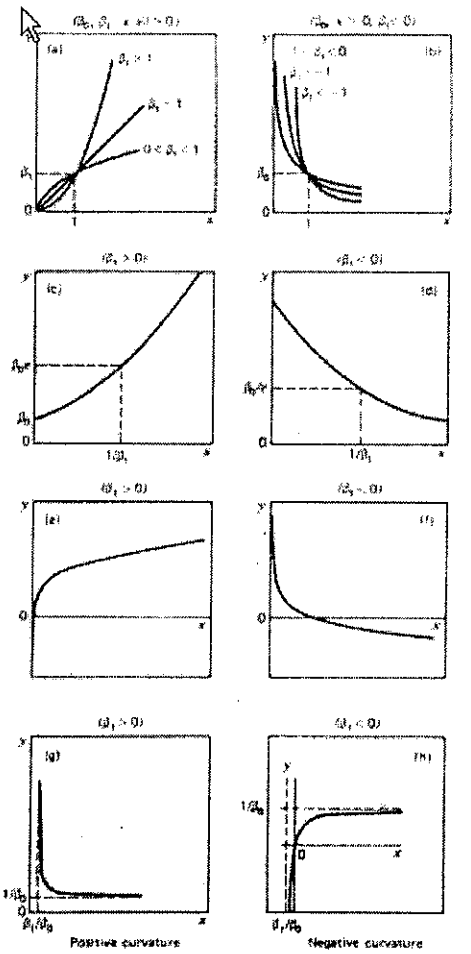
$$b_0 \sum_{i=1}^{n} w_i X_i + b_1 \sum_{i=1}^{n} w_i X_i^2 = \sum_{i=1}^{n} w_i X_i Y_i \tag{4.27}$$

แก้สมการปกติทั้งสองสมการข้างต้น จะได้ตัวประมาณกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (Weighted Least

Square Estimator: WLS) ซึ่งแทนด้วย $\mathbf{b}_{WLS}$

จะเห็นได้ว่าการหาตัวประมาณกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักจะต้องทราบค่าน้ำหนัก w_i ซึ่งในบางกรณีสามารถกำหนดได้ง่าย เช่น ถ้า $V(Y_i) = \sigma_i^2 = c_i^2\sigma^2$ เมื่อ c_i^2 เป็นค่าคงที่ที่ทราบค่า ($c_i = 1$ ได้ตัวประมาณกำลังสองน้อยที่สุด) จะได้ว่า $w_i = 1/c_i^2$ หากค่าสังเกต Y_i เกิดจากการเฉลี่ยค่าสังเกต n_i ค่าที่จุด X_i และถ้าค่าสังเกตเดิมมีความแปรปรวนคงที่ σ^2 จะได้ว่า $V(Y_i) = V(\epsilon_i) = \sigma^2/n_i$ ดังนั้น $w_i = n_i$ หาก $V(Y_i)$ เป็นฟังก์ชันของตัวแปรอิสระ $V(Y_i) = V(\epsilon_i) = \sigma^2 X_i$ จะได้ว่า $w_i = 1/X_i$ หากความผันแปรหลักเกิดจากความคลาดเคลื่อนในการวัดและใช้เครื่องมือที่ไม่แม่นยำแต่ทราบค่าที่แน่นอน อาจเลือก w_i ให้เป็นสัดส่วนผกผันกับความผันแปรของ Y_i ที่เกิดจากการวัด แต่ในบางกรณีที่ไม่ทราบค่า w_i จะต้องมีการประมาณค่า ซึ่งมีวิธีในการประมาณหลายวิธีด้วยกัน สามารถศึกษาเพิ่มเติมได้จาก (?)

4.5 การตรวจสอบความสัมพันธ์เชิงเส้นตรง



รูปที่ 4.10: กราฟแสดงลักษณะฟังก์ชันที่สามารถแปลงให้เป็นฟังก์ชันเชิงเส้นตรงได้ (จาก Montgomery และ Peck (1992))

ข้อกำหนดเกี่ยวกับลักษณะความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรตามและตัวแปรอิสระจัดเป็นจุดเริ่มต้นของการวิเคราะห์การถดถอย บางกรณีอาจพบว่ารูปแบบความสัมพันธ์เชิงเส้นตรง (Nonlinearity) นั้นไม่เหมาะสม ซึ่งการตรวจสอบความเหมาะสมของสมการเชิงเส้นตรงสามารถทำได้โดยใช้การทดสอบ Lack of fit สร้างแผนภาพการกระจาย หรือสร้างกราฟของความคลาดเคลื่อน นอกจากนี้ทฤษฎีและประสบการณ์ในอดีตอาจชี้ว่ารูปแบบความสัมพันธ์ไม่ใช่เชิงเส้นตรง และในบางครั้งการแปลงข้อมูลที่เหมาะสมอาจเปลี่ยนโมเดลที่ไม่ใช่เส้นตรงให้เป็นโมเดลเชิงเส้นตรงได้ โดยจะเรียกโมเดลที่มีลักษณะดังกล่าวนี้ว่า *Intrinsically* หรือ *transformably linear*

รูปที่ 4.10 แสดงลักษณะของฟังก์ชันที่สามารถแปลงให้เป็นฟังก์ชันเชิงเส้นตรง (Linearizable function) ได้ และวิธีการแปลงข้อมูลที่สอดคล้องกันเพื่อให้ฟังก์ชันมีลักษณะเป็นเส้นตรงแสดงในตารางที่ 4.13 เช่น พิจารณาฟังก์ชันเอกซ์โปเนนเชียล

$$Y = \beta_0 e^{\beta_1 X} \epsilon$$

ซึ่งสามารถแปลงให้เป็นฟังก์ชันเชิงเส้นตรงได้ดังนี้

$$\ln Y = \ln \beta_0 + \beta_1 X + \ln \epsilon$$

หรือ

$$Y' = \beta_0{}' + \beta_1 X + \epsilon'$$

โดยการแปลงข้อมูลดังกล่าวกำหนดว่า $\epsilon' = \ln \epsilon$ มีการแจกแจงแบบปกติและเป็นอิสระกัน ด้วยค่าเฉลี่ยเป็น 0 และความแปรปรวน σ^2 ซึ่งแสดงว่าความคลาดเคลื่อนในโมเดลเดิมมีการแจกแจงแบบล็อกนอร์มอล (Log-normal distribution) นั่นเอง ดังนั้นหลังจากที่มีการแปลงข้อมูล ควรที่จะพิจารณาว่าข้อกำหนดของการวิเคราะห์เป็นจริงหรือไม่ นอกจากนี้ตัวประมาณกำลังสองน้อยที่สุดที่ได้จากโมเดลที่มีการแปลงให้เป็นเส้นตรงจะมีคุณสมบัติสอดคล้องกับข้อมูลที่แปลงค่าแล้ว ไม่ใช่ข้อมูลเดิม

รูปที่	ฟังก์ชันที่แปลงเป็นเส้นตรงได้	วิธีแปลงข้อมูล	รูปแบบเชิงเส้นตรง
4.10 a, b	$Y = \beta_0 X^{\beta_1}$	$Y' = \log Y,\ X' = \log X$	$Y' = \log \beta_0 + \beta_1 X'$
4.10 c, d	$Y = \beta_0 e^{\beta_1 X}$	$Y' = \ln Y$	$Y' = \ln \beta_0 + \beta_1 X$
4.10 e, f	$Y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y' = \beta_0 + \beta_1 X'$
4.10 g, h	$Y = \dfrac{X}{\beta_0 X - \beta_1}$	$Y' = \dfrac{1}{Y}, X' = \dfrac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$

ตารางที่ 4.13: แสดงฟังก์ชันที่แปลงเป็นเส้นตรงได้และวิธีการแปลงข้อมูลที่สอดคล้องกัน

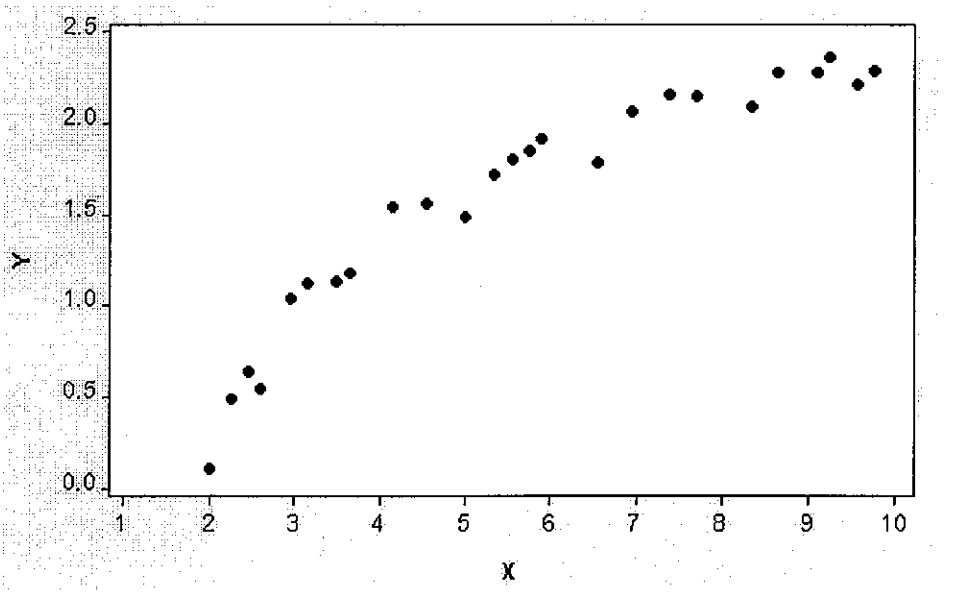
ตัวอย่างที่ 4.8 ข้อมูลต่อไปนี้แสดงระดับความเข้มข้นของโซเดียมไฮดรอกไซด์ (X) และปริมาณน้ำตาล (Y) ที่ได้

X	4.55	5.55	2.95	2.25	9.55	9.25	9.10	2.60	7.70	5.75
Y	1.567	1.807	1.042	0.485	2.221	2.371	2.279	0.543	2.151	1.851
X	2.45	5.90	4.15	5.35	6.95	3.15	7.40	8.35	6.55	5.00
Y	0.638	1.915	1.547	1.722	2.073	1.122	2.164	2.097	1.785	1.486
X	8.65	9.75	3.65	3.50	2.00					
Y	2.288	2.295	1.179	1.129	0.108					

ตารางที่ 4.14: ข้อมูลแสดงระดับความเข้มข้นของโซเดียมไฮดรอกไซด์ (X) และปริมาณน้ำตาล (Y)

จงสร้างสมการแสดงความสัมพันธ์ที่เหมาะสมของข้อมูลชุดนี้

วิธีทำ



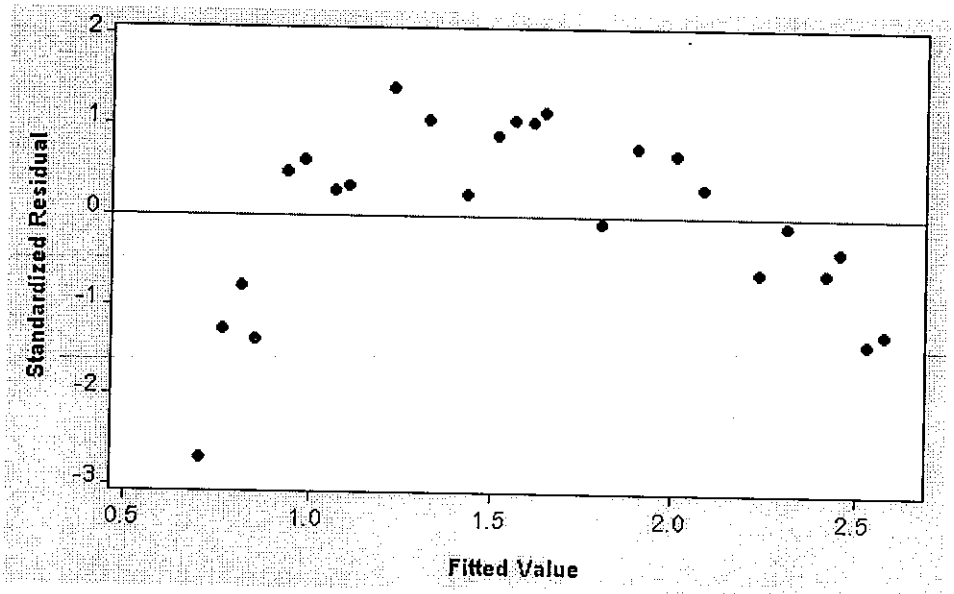
รูปที่ 4.11: แผนภาพการกระจายระหว่างระดับความเข้มข้นของโซเดียมไฮดรอกไซด์และปริมาณน้ำตาล

จากแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างระดับความเข้มข้นของโซเดียมไฮดรอกไซด์และปริมาณน้ำตาลในรูปที่ 4.11 พบว่าลักษณะความสัมพันธ์อาจไม่ใช่เชิงเส้นตรง แต่หากสร้างสมการถดถอยเชิงเส้นตรงกับข้อมูลชุดนี้จะได้ว่า

$$\hat{Y} = 0.2244 + 0.2412X$$

เมื่อพิจารณากราฟของความคลาดเคลื่อนที่ได้จากสมการข้างต้น ดังแสดงในรูปที่ 4.12 จะเห็นได้ว่ารูปแบบโมเดลถดถอยเชิงเส้นตรงไม่เหมาะสมกับข้อมูลดังกล่าว ดังนั้นจึงควรที่จะพิจารณารูปแบบอื่นของโมเดล เช่น โมเดลควอดราติก ซึ่งมีรูปแบบดังนี้

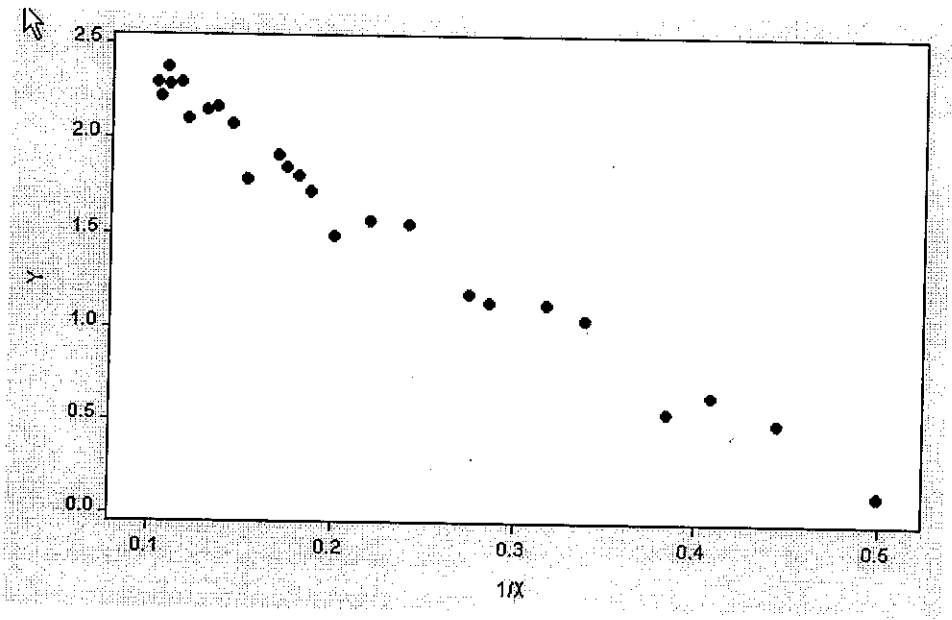
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$



รูปที่ 4.12: กราฟระหว่างคะแนนมาตรฐานของความคลาดเคลื่อนและค่าพยากรณ์

จากรูปที่ 4.11 พบว่า เมื่อความเข้มข้นของโซเดียมไฮดรอกไซด์เพิ่มขึ้น ปริมาณน้ำตาลจะสูงขึ้นจนกระทั่งถึงขีด-จำกัดบน (Upper limit) แล้วจึงมีค่าคงที่ หากใช้โมเดลควอดราติกอธิบายความสัมพันธ์ของข้อมูลดังกล่าว พบว่าในท้ายที่สุดโมเดลควอดราติกจะทำให้ปริมาณน้ำตาลลดลง เมื่อความเข้มข้นของโซเดียมไฮดรอกไซด์เพิ่มขึ้นเรื่อย ๆ จึงอาจไม่ใช่รูปแบบโมเดลที่เหมาะสมกับข้อมูลชุดนี้ ดังนั้นโมเดลที่น่าจะมีลักษณะที่เหมาะสมมากกว่าคือ

$$Y = \beta_0 + \beta_1 \left(\frac{1}{X}\right) + \epsilon$$

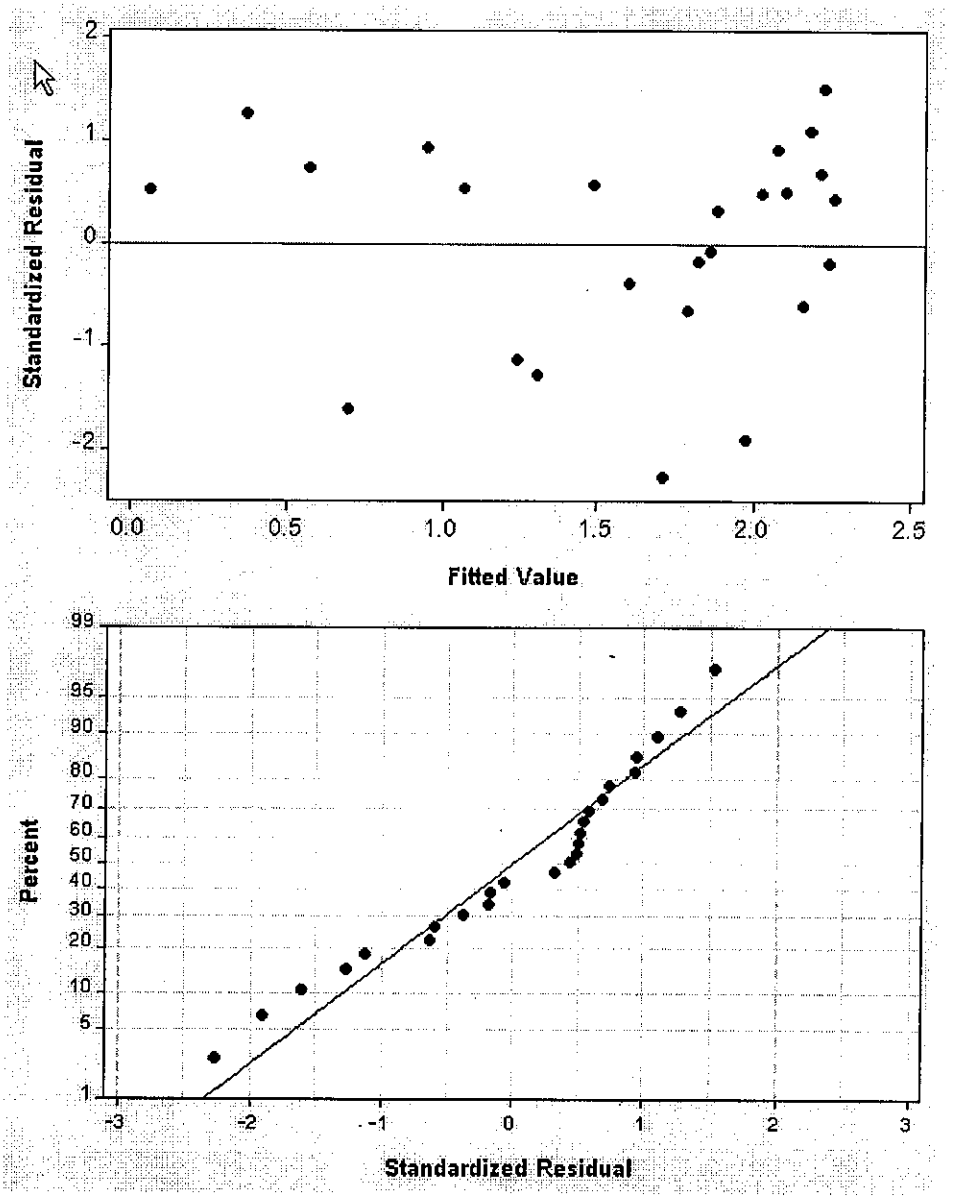


รูปที่ 4.13: แสดงกราฟระหว่าง Y กับ $X' = 1/X$

รูปที่ 4.13 แสดงแผนภาพการกระจายระหว่างตัวแปรตาม Y กับตัวแปรอิสระที่มีการแปลงค่า $X' = 1/X$ จะเห็นได้ว่าความสัมพันธ์ระหว่างตัวแปรทั้งสองมีลักษณะเป็นเส้นตรง ซึ่งชี้ว่าการแปลงข้อมูลโดยใช้ $X' = 1/X$ มีความเหมาะสม สร้างสมการถดถอยสำหรับข้อมูลที่มีการแปลงค่าแล้วได้ดังนี้

$$\hat{Y} = 2.8167 - 5.5079X'$$

โดยที่ $R^2 = 0.976$ และ $MSE = 0.0105$



รูปที่ 4.14: กราฟระหว่างคะแนนมาตรฐานของความคลาดเคลื่อนและค่าพยากรณ์ (บน) และ Normal probability plot (ล่าง) ของข้อมูลที่แปลงค่าแล้ว

พิจารณารูปที่ 4.14 (บน) จะเห็นได้ว่าความคลาดเคลื่อนที่ได้จากโมเดลที่มีการแปลงค่าแล้วมีการกระจายอย่างสุ่ม ไม่มีรูปแบบที่แน่นอน และจาก Normal probability plot ในรูปที่ 4.14 (ล่าง) พบว่าข้อมูลชุดนี้มีการแจกแจงที่มีหางหนากว่าการแจกแจงแบบปกติเล็กน้อย แต่เนื่องจากไม่มีสัญญาณที่รุนแรงชี้ว่าข้อมูลไม่เป็นไปตามข้อกำหนดของการวิเคราะห์ ดังนั้นอาจถือว่าโมเดลที่ได้จากข้อมูลที่มีการแปลงค่าแล้วมีความเหมาะสม

4.6 การตรวจสอบความสัมพันธ์ของความคลาดเคลื่อน

ข้อกำหนดเบื้องต้นของการวิเคราะห์การถดถอยประกอบด้วย ความคลาดเคลื่อนมีค่าเฉลี่ยเป็น 0 ความแปรปรวนคงที่ และเป็นอิสระกัน หากต้องการสร้างช่วงความเชื่อมั่นและทดสอบสมมติฐาน จะเพิ่มข้อกำหนดเกี่ยวกับการแจกแจงแบบปกติของความคลาดเคลื่อนเข้าไปด้วย และได้ว่า $\epsilon_i \sim NID(0, \sigma^2)$ ในบางครั้งการนำสมการถดถอยไปประยุกต์ใช้พบว่า ตัวแปรอิสระและตัวแปรตามเรียงตามลำดับเวลา ซึ่งมักเกิดกับข้อมูลทางด้านเศรษฐศาสตร์ ธุรกิจ และวิศวกรรมบางสาขา โดยจะเรียกข้อมูลลักษณะนี้ว่า *ข้อมูลอนุกรมเวลา (Time series data)* ซึ่งความคลาดเคลื่อนในข้อมูลอนุกรมเวลามักมีความสัมพันธ์ต่อกัน (Serial correlation) นั่นคือ $E(\epsilon_i \epsilon_j) \neq 0$, เมื่อ $i \neq j$ หรือเกิดปัญหา Autocorrelation นั่นเอง ซึ่งสาเหตุของปัญหา Autocorrelation ในการวิเคราะห์การถดถอยที่เกี่ยวข้องกับข้อมูลอนุกรมเวลามักเกิดขึ้นเนื่องจากขาดตัวแปรอิสระที่สำคัญบางตัวในโมเดล เช่น การศึกษาความสัมพันธ์ระหว่างราคาขายบ้านกับขนาดของตัวบ้าน พบว่าอัตราการเพิ่มของประชากรหรือความหนาแน่นของประชากรต่อพื้นที่อาจมีผลกระทบต่อราคาขายบ้านเช่นกัน หากโมเดลถดถอยไม่ได้รวมตัวแปรดังกล่าวไว้ อาจก่อให้เกิดปัญหา Positive autocorrelation ได้ เนื่องจากขนาดของประชากรมีความสัมพันธ์เชิงบวกกับราคาขาย

เมื่อเกิดปัญหา Autocorrelation จะส่งผลกระทบต่อการประมาณพารามิเตอร์ในโมเดลถดถอยด้วยวิธีกำลังสองน้อยที่สุดดังนี้

- ค่าสัมประสิทธิ์ถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุดยังคงเป็นตัวประมาณที่ไม่เอนเอียง (Unbiased) แต่ไม่มีความแปรปรวนต่ำที่สุด ทำให้ได้ตัวประมาณที่ไม่มีประสิทธิภาพ (Inefficient)
- เมื่อเกิดปัญหา Positive autocorrelation พบว่า MSE ให้ค่าประมาณของ σ^2 ที่ต่ำกว่าความเป็นจริง (Underestimate) ส่งผลให้ค่าความคลาดเคลื่อนมาตรฐานของตัวประมาณมีค่าต่ำเกินไป ทำให้ได้ช่วงความเชื่อมั่นแคบกว่าที่ควรจะเป็น และในการทดสอบสมมติฐานเกี่ยวกับค่าสัมประสิทธิ์ถดถอย พบว่ามีตัวแปรอิสระตั้งแต่หนึ่งตัวขึ้นไปมีอิทธิพลต่อโมเดลอย่างมีนัยสำคัญ ทั้งที่จริงตัวแปรเหล่านี้อาจไม่มีอิทธิพลต่อโมเดลที่พิจารณา
- การสร้างช่วงความเชื่อมั่นและทดสอบสมมติฐานโดยใช้สถิติทดสอบ t และ F ไม่เหมาะสม

4.6.1 การตรวจสอบปัญหา Autocorrelation

ดังได้กล่าวแล้วว่ากราฟของความคลาดเคลื่อนกับลำดับเวลามีประโยชน์ในการตรวจสอบปัญหา Autocorrelation หากความคลาดเคลื่อนมีเครื่องหมายเหมือนกันเป็นกลุ่ม ๆ นั่นคือมีการเปลี่ยนแปลงเครื่องหมายของความคลาดเคลื่อนน้อยเกินไป แสดงว่าเกิดปัญหา Positive autocorrelation ในทางตรงข้ามหากความคลาดเคลื่อนมีการเปลี่ยนแปลงเครื่องหมายบ่อยครั้งเกินไป แสดงว่าเกิดปัญหา Negative autocorrelation

วิธีตรวจสอบการเกิด Autocorrelation ได้พัฒนาโดยนักสถิติหลายท่าน ซึ่งสถิติทดสอบของ Durbin และ Watson เป็นสถิติที่ใช้กันอย่างแพร่หลาย โดยกำหนดว่าข้อมูลมีระยะห่างของช่วงเวลาเท่า ๆ กัน และความคลาดเคลื่อนในโมเดลถดถอยอยู่ในกระบวนการ Autoregressive อันดับที่ 1 หรือนิยมเรียกย่อ ๆ ว่า AR(1) ในที่นี้จะกล่าวถึงสถิติทดสอบของ Durbin และ Watson โดยพิจารณาเฉพาะโมเดลถดถอยเชิงเส้นตรงอย่างง่ายเท่านั้น

พิจารณาโมเดลของความคลาดเคลื่อน AR(1)

$$\epsilon_t = \rho\epsilon_{t-1} + a_t \tag{4.28}$$

เมื่อ

ϵ_t แทน ความคลาดเคลื่อนในโมเดล ณ เวลา t

a_t แทน ตัวรบกวนสุ่ม (Random disturbance) โดยที่ $a_t \sim NID(0, \sigma_a^2)$

ρ แทน Autocorrelation parameter ซึ่งเป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง ϵ_t กับ ϵ_{t-1} โดยที่ $|\rho| < 1$ ถ้า $\rho > 0$ เกิด Positive autocorrelation และถ้า $\rho < 0$ เกิด Negative autocorrelation

ดังนั้นโมเดลถดถอยเชิงเส้นตรงอย่างง่ายที่รวมเอารูปแบบ AR(1) ของความคลาดเคลื่อน สามารถแสดงได้ดังนี้

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + a_t \end{aligned} \tag{4.29}$$

เมื่อ

Y_t แทน ค่าสังเกตของตัวแปรตาม ณ เวลา t

X_t แทน ค่าสังเกตของตัวแปรอิสระ ณ เวลา t

จะเห็นได้ว่าสมการข้างต้นมีรูปแบบคล้ายกับสมการถดถอยเชิงเส้นตรงอย่างง่ายทั่วไป เพียงแต่ความคลาดเคลื่อน ณ เวลาปัจจุบันเป็นฟังก์ชันของความคลาดเคลื่อน ณ เวลาก่อนหน้า สำหรับโมเดลถดถอยเชิงเส้นตรงแบบพหุที่รวมเอารูปแบบ AR(1) ของความคลาดเคลื่อนไว้ สามารถแสดงได้ในลักษณะที่คล้ายกันดังนี้

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \ldots + \beta_k X_{tk} + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + a_t \end{aligned} \tag{4.30}$$

โดยความคลาดเคลื่อนที่อยู่ในกระบวนการ AR(1) มีคุณสมบัติดังนี้ เมื่อแทนค่า ϵ_{t-1} ในสมการที่ (4.28) จะได้ว่า

$$\begin{aligned} \epsilon_t &= \rho\epsilon_{t-1} + a_t \\ &= \rho\left(\rho\epsilon_{t-2} + a_{t-1}\right) + a_t \\ &= \rho^2\epsilon_{t-2} + \rho a_{t-1} + a_t \end{aligned}$$

เมื่อแทนค่า ϵ_{t-2} เข้าในสมการข้างต้น จะได้ว่า

$$\begin{aligned} \epsilon_t &= \rho^2\left(\rho\epsilon_{t-3} + a_{t-2}\right) + \rho a_{t-1} + a_t \\ &= \rho^3\epsilon_{t-3} + \rho^2 a_{t-2} + \rho a_{t-1} + a_t \end{aligned}$$

แทนค่า $\epsilon_{t-4}, \epsilon_{t-5}, \ldots$ ต่อเนื่องกันไป จะได้ว่า

$$\epsilon_t = \sum_{u=0}^{\infty} \rho^u a_{t-u} \tag{4.31}$$

จะเห็นได้ว่าความคลาดเคลื่อน ณ เวลา t เป็นฟังก์ชันเชิงเส้นของตัวแปรสุ่ม a_t ที่เกิดขึ้นก่อนหน้า โดย $a_t \sim NID(0, \sigma_a^2)$ และ $|\rho| < 1$ นั่นคือ เมื่อเวลายิ่งห่างออกไป $|\rho|$ จะมีค่าลดลง ทำให้น้ำหนักบนตัวแปรสุ่ม a_{t-u} มีค่าลดลงตามไปด้วย นอกจากนี้ยังสามารถแสดงได้ว่า

$$\begin{aligned} E(\epsilon_t) &= E\left(\sum_{u=0}^{\infty} \rho^u a_{t-u}\right) \\ &= \sum_{u=0}^{\infty} \rho^u E(a_{t-u}) \\ &= 0 \end{aligned} \tag{4.32}$$

เนื่องจาก $E(a_{t-u}) = 0,\ u = 0, 1, \ldots$

$$\begin{aligned} V(\epsilon_t) &= V\left(\sum_{u=0}^{\infty} \rho^u a_{t-u}\right) \\ &= \sum_{u=0}^{\infty} \rho^{2u} V(a_{t-u}) \\ &= \sigma_a^2 \sum_{u=0}^{\infty} \rho^{2u} \\ &= \sigma_a^2 \cdot \left(\frac{1}{1-\rho^2}\right) \\ &= \frac{\sigma_a^2}{1-\rho^2} \end{aligned} \tag{4.33}$$

พิจารณาความแปรปรวนร่วมระหว่างความความคลาดเคลื่อน ณ เวลา t กับความคลาดเคลื่อน ณ เวลาก่อนหน้า 1 ช่วงเวลา

$$\begin{aligned} Cov(\epsilon_t, \epsilon_{t-1}) &= E(\epsilon_t\, \epsilon_{t-1}), \qquad \text{เนื่องจาก} \quad E(\epsilon_t) = E(\epsilon_{t-1}) = 0 \\ &= E\left(\left\{a_t + \rho a_{t-1} + \rho^2 a_{t-2} + \cdots\right\} \cdot \left\{a_{t-1} + \rho a_{t-2} + \rho^2 a_{t-3} + \cdots\right\}\right) \\ &= E\left(\left\{a_t + \rho[a_{t-1} + \rho a_{t-2} + \cdots]\right\} \cdot \left\{a_{t-1} + \rho a_{t-2} + \rho^2 a_{t-3} + \cdots\right\}\right) \\ &= E\left(a_t \left\{a_{t-1} + \rho a_{t-2} + \cdots\right\}\right) + E\left(\rho \left\{a_{t-1} + \rho a_{t-2} + \rho^2 a_{t-3} + \cdots\right\}^2\right) \\ &= E(\rho \epsilon_{t-1}^2), \qquad \text{เนื่องจาก} \quad E(a_t a_{t-u}) = 0,\ \forall u \neq 0 \\ &= \rho E(\epsilon_{t-1}^2) \\ &= \rho V(\epsilon) \\ &= \rho\left(\frac{\sigma^2}{1-\rho^2}\right) \end{aligned} \tag{4.34}$$

และได้ค่าสัมประสิทธิ์สหสัมพันธ์เป็น

$$\begin{aligned} Corr(\epsilon_t, \epsilon_{t-1}) &= \frac{Cov(\epsilon_t, \epsilon_{t-1})}{\sqrt{V(\epsilon_t)} \cdot \sqrt{V(\epsilon_{t-1})}} \\ &= \frac{\rho\left(\frac{\sigma_a^2}{1-\rho^2}\right)}{\sqrt{\frac{\sigma_a^2}{1-\rho^2}} \cdot \sqrt{\frac{\sigma_a^2}{1-\rho^2}}} \\ &= \rho \end{aligned} \tag{4.35}$$

ดังนั้นความแปรปรวนร่วมและสัมประสิทธิ์สหสัมพันธ์ระหว่างความคลาดเคลื่อนที่ห่างกัน u ช่วงเวลา สามารถเขียนให้อยู่ในรูปสูตรทั่วไปได้ดังนี้

$$Cov(\epsilon_t, \epsilon_{t-u}) = \rho^{|u|}\left(\frac{\sigma_a^2}{1-\rho^2}\right), \qquad u \neq 0 \tag{4.36}$$

$$Corr(\epsilon_t, \epsilon_{t-u}) = \rho^{|u|} \tag{4.37}$$

เนื่องจาก $|\rho| < 0$ จะเห็นได้ว่าเมื่อช่วงเวลายิ่งห่างกันมากขึ้น ความคลาดเคลื่อนจะมีความสัมพันธ์กันน้อยลง และไม่มีความสัมพันธ์เชิงเส้นตรงต่อกันเมื่อ $\rho = 0$ เนื่องจาก $\epsilon_t = a_t$ และตัวรบกวนสุ่ม a_t เป็นอิสระกัน ดังนั้นการตรวจสอบ Autocorrelation ทำได้โดยทดสอบว่า $\rho = 0$ หรือไม่ ซึ่งการประยุกต์ทางด้านเศรษฐศาสตร์และธุรกิจมักเกิดปัญหา Positive autocorrelation และสามารถทดสอบได้ดังนี้
กำหนดสมมติฐานของการทดสอบ

$$H_{01} : \rho = 0 \qquad vs. \qquad H_{11} : \rho > 0$$

คำนวณสถิติทดสอบ Durbin-Watson

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{4.38}$$

เมื่อ e_t แทน ความคลาดเคลื่อนของสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด โดยที่ $e_t = Y_t - \hat{Y}_t$, $t = 1, 2, \ldots, n$ และ n แทน จำนวนค่าสังเกตทั้งหมด

การคำนวณค่าวิกฤติที่แท้จริงของสถิติทดสอบ Durbin-Watson ทำได้ยาก Durbin และ Watson จึงได้กำหนดขอบเขตล่าง d_L และขอบเขตบน d_U ของสถิติทดสอบดังกล่าว หากค่าที่คำนวณตกอยู่นอกขอบเขตที่กำหนดจะนำไปสู่การตัดสินใจดังนี้

ถ้า $d < d_L$ จะปฏิเสธ H_0

ถ้า $d > d_U$ จะยอมรับ H_0

ถ้า $d_L \leq d \leq d_U$ สรุปไม่ได้

โดยค่าน้อย ๆ ของ d จะนำไปสู่การปฏิเสธ $H_0 : \rho = 0$ นั่นคือ เกิดปัญหา Positive autocorrelation เนื่องจากความคลาดเคลื่อนที่อยู่ติดกัน e_t และ e_{t-1} มีแนวโน้มที่จะมีค่าใกล้เคียงกันเมื่อมีความสัมพันธ์ไป

ทางเดียวกัน ดังนั้นผลต่างของความคลาดเคลื่อนที่อยู่ติดกัน $e_t - e_{t-1}$ จึงมีค่าน้อยเมื่อ $\rho > 0$ และทำให้สถิติทดสอบ Durbin-Watson มีค่าน้อยตามไปด้วย ตารางที่ 5 (ภาคผนวก) แสดงค่าวิกฤติของสถิติทดสอบ Durbin-Watson ที่ระดับนัยสำคัญและขนาดตัวอย่างที่แตกต่างกัน

หากต้องการทดสอบ Negative autocorrelation ซึ่งมักเกิดขึ้นไม่บ่อยนัก ก็สามารถทำได้ในลักษณะที่คล้ายคลึงกัน โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_{02} : \rho = 0 \qquad vs. \qquad H_{12} : \rho < 0$$

คำนวณสถิติทดสอบจาก $4 - d$

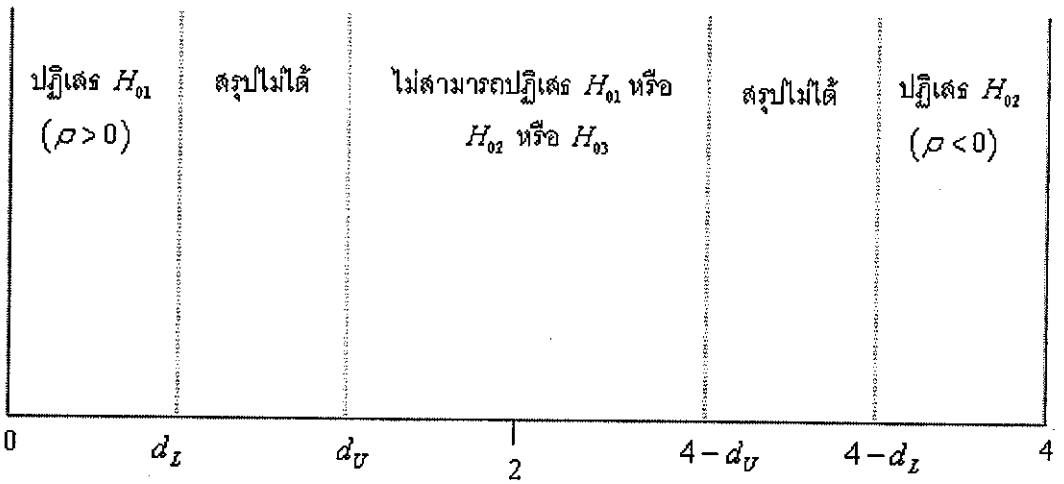
เมื่อ d แทน สถิติทดสอบ Durbin-Watson ที่คำนวณได้จากสมการ (4.38)

เกณฑ์ในการตัดสินใจยังคงใช้เกณฑ์เดียวกับการทดสอบ Positive autocorrelation ดังนี้

ถ้า $4 - d < d_L$ จะปฏิเสธ H_0

ถ้า $4 - d > d_U$ จะยอมรับ H_0

ถ้า $d_L \leq 4 - d \leq d_U$ สรุปผลไม่ได้



รูปที่ 4.15: เกณฑ์การตัดสินใจของสถิติทดสอบ Durbin-Watson

สำหรับการทดสอบสองทางจะกำหนดสมมติฐานของการทดสอบเป็น

$$H_{03} : \rho = 0 \qquad vs. \qquad H_{13} : \rho \neq 0$$

โดยทำการทดสอบแบบทางเดียวพร้อม ๆ กัน ซึ่งทำให้ความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของการทดสอบสองทางมีค่าเท่ากับ 2α เมื่อ α เป็นความคลาดเคลื่อนประเภทที่ 1 ของการทดสอบทางเดียว ซึ่ง

เกณฑ์การตัดสินใจของสถิติ Durbin-Watson สามารถสรุปได้ดังแสดงในรูปที่ 4.15

หมายเหตุ

- เมื่อใช้สถิติทดสอบของ Durbin-Watson และไม่สามารถสรุปผลได้ โดยทั่วไปแสดงว่าโมเดลต้องการจำนวนค่าสังเกตมากขึ้น แต่สำหรับข้อมูลอนุกรมเวลาแล้ว อาจทำไม่ได้ เพราะค่าสังเกตที่เพิ่มขึ้น อาจหมายถึงค่าสังเกตในอนาคต ซึ่งต้องใช้เวลาในการเก็บข้อมูล ทำให้การวิเคราะห์ล่าช้า ทางเลือกอีกทางก็คือ พิจารณาผลการวิเคราะห์ที่ไม่สามารถสรุปผลได้นั้นราวกับว่าเกิดปัญหา Autocorrelation และทำการแก้ไข หากผลการวิเคราะห์หลังจากที่ได้แก้ไขแล้วแตกต่างจากเดิมไม่มากนัก แสดงว่าความคลาดเคลื่อนเป็นอิสระกัน ซึ่งสอดคล้องกับข้อกำหนดของการวิเคราะห์ แต่หากผลการวิเคราะห์แตกต่างจากเดิมมาก เช่น ค่าความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์ถดถอยมีค่าสูงมาก เป็นต้น แสดงว่าเกิดปัญหา Autocorrelation ดังนั้นผลการวิเคราะห์ที่ได้หลังจากแก้ไขแล้วจะมีประโยชน์มากกว่าผลการวิเคราะห์เดิม
- สถิติทดสอบ Durbin-Watson ไม่มีความแกร่ง (Robust) เมื่อกำหนดรูปแบบโมเดลไม่ถูกต้อง (Model misspecificaiton) นั่นคือ สถิติทดสอบ Durbin-Watson อาจไม่พบปัญหา Autocorrelation สำหรับโมเดลที่มีความคลาดเคลื่อนแบบ AR(2)
- การตรวจสอบ Autocorrelation นอกจากใช้สถิติทดสอบ Durbin-Watson แล้ว ยังมีสถิติอื่นที่สามารถใช้ตรวจสอบได้เช่นกัน เช่น สถิติทดสอบของ Theil และ Nagar

ตัวอย่างที่ 4.9 ข้อมูลในตารางที่ 4.15 แสดงยอดขาย (Y) และค่าใช้จ่ายในการโฆษณา (X) รายไตรมาสระหว่างปีพ.ศ. 2536 ถึง 2540 จงตรวจสอบว่าข้อมูลชุดนี้เกิดปัญหา Positive autocorrelation หรือไม่

วิธีทำ สร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายของข้อมูลในตารางที่ 4.15 ได้ดังนี้

$$\hat{Y}_t = \underset{(0.2141)}{-1.4548} + \underset{(0.0014)}{0.1763} X_t$$

$MSE = 0.0074$

โดยตัวเลขในวงเล็บใต้ค่าสัมประสิทธิ์ถดถอยแทนค่าความคลาดเคลื่อนมาตรฐานของตัวประมาณที่สอดคล้องกัน เมื่อพิจารณารูปที่ 4.16 จะเห็นได้ว่าความคลาดเคลื่อนของข้อมูลที่เวลาใกล้เคียงกัน มีแนวโน้มไปทางเดียวกัน ซึ่งชี้ว่าข้อมูลชุดนี้อาจเกิดปัญหา Positive autocorrelation จากนั้นทำการตรวจสอบโดยใช้สถิติ Durbin-Watson โดยกำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \rho = 0 \qquad vs. \qquad H_1 : \rho > 0$$

ปีพ.ศ.	ไตรมาสที่	Y_t	X_t	e_t	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	e_t^2
2536	1	20.96	127.3	-0.0261	-	-	0.00068
	2	21.40	130.0	-0.0620	-0.0360	0.001293	0.00385
	3	21.96	132.7	0.0220	0.0840	0.007062	0.00048
	4	21.52	129.4	0.1638	0.1417	0.020088	0.02682
2537	1	22.39	135.0	0.0466	-0.1172	0.013732	0.00217
	2	22.76	137.1	0.0464	-0.0002	0.000000	0.00215
	3	23.48	141.2	0.0436	-0.0028	0.000008	0.00190
	4	23.66	142.8	-0.0584	-0.1021	0.010415	0.00341
2538	1	24.10	145.5	-0.0944	-0.0360	0.001293	0.00891
	2	24.01	145.3	-0.1491	-0.0547	0.002997	0.02224
	3	24.54	148.3	-0.1480	0.0012	0.000001	0.02190
	4	24.30	146.4	-0.0531	0.0949	0.009013	0.00281
2539	1	25.00	150.2	-0.0229	0.0301	0.000908	0.00053
	2	25.64	153.1	0.1059	0.1288	0.016584	0.01120
	3	26.36	157.3	0.0855	-0.0204	0.000416	0.00730
	4	26.98	160.7	0.1061	0.0206	0.000426	0.01126
2540	1	27.52	164.2	0.0291	-0.0770	0.005927	0.00085
	2	27.78	165.6	0.0423	0.0132	0.000174	0.00179
	3	28.24	168.7	-0.0442	-0.0865	0.007478	0.00195
	4	28.78	171.7	-0.0330	0.0112	0.000124	0.00109
	รวม	491.38	2952.50	0.0000		0.097940	0.13330

ตารางที่ 4.15: ข้อมูลยอดขาย (Y_t) และค่าใช้จ่ายในการโฆษณา (X_t) ระหว่างปีพ.ศ. 2536 ถึง 2540

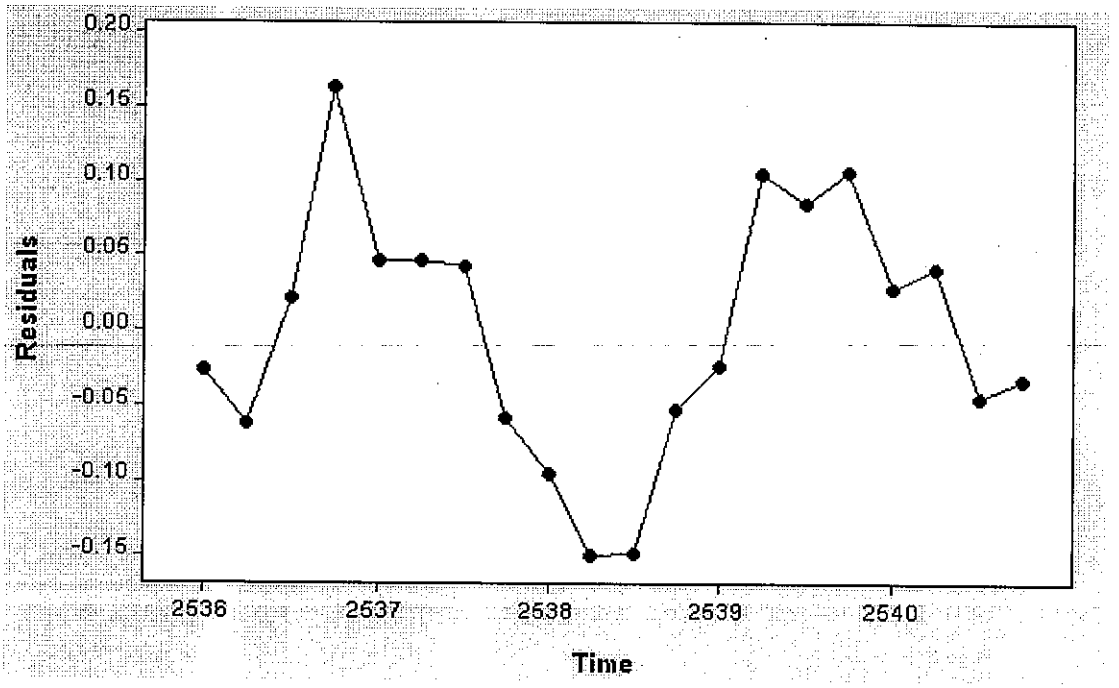
คำนวณสถิติทดสอบ Durbin-Watson

$$\begin{aligned} d &= \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \\ &= \frac{0.097940}{0.13330} \\ &= 0.735 \end{aligned}$$

จากตารางค่าวิกฤติของ Durbin-Watson ที่ $\alpha = 0.05$, $k = 1$ และ $n = 20$ ได้ค่า $d_L = 1.20$ และ $d_U = 1.41$ เนื่องจาก $d = 0.735 < d_L$ จึงปฏิเสธ H_0 นั่นคือ ความคลาดเคลื่อนเกิด Positive autocorrelation

ถึงแม้ว่าสถิติทดสอบ Durbin-Watson มีประโยชน์อย่างมาก แต่ก็ยังมีข้อจำกัดเกี่ยวกับรูปแบบของความคลาดเคลื่อนว่าต้องอยู่ในกระบวนการ AR(1) จึงอาจไม่สามารถตรวจสอบ Autocorrelation เมื่อความคลาดเคลื่อนมีรูปแบบอื่น นอกจากนี้ข้อมูลอนุกรมเวลายังอาจเกี่ยวข้องกับตัวแปรตามที่มีลักษณะเป็น Lagged variable เช่น

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \epsilon$$



รูปที่ 4.16: กราฟระหว่างความคลาดเคลื่อนและลำดับเวลาของข้อมูลในตารางที่ 4.15

หากประมาณพารามิเตอร์ในโมเดลข้างต้นด้วยวิธีกำลังสองน้อยที่สุด ตัวประมาณที่ได้จะเป็นตัวประมาณที่เอนเอียง (Bias) และไม่คงเส้นคงวา (Consistent) นอกจากนี้ยังพบว่าการตรวจสอบ Autocorrelation ด้วยสถิติ Durbin-Watson สำหรับโมเดลที่มี Lagged variable ไม่เหมาะสมอีกต่อไป

4.6.2 การแก้ไขปัญหา Autocorrelation

วิธีแก้ไขโดยทั่วไปเมื่อเกิดปัญหา Autocorrelation ทำได้ดังนี้ หากทราบว่าปัญหา Autocorrelation นั้นเกิดจากการละตัวแปรอิสระที่สำคัญบางตัวในโมเดล สามารถแก้ไขได้โดยระบุตัวแปรดังกล่าวและเพิ่มเข้ามาในโมเดล ซึ่งปัญหา Autocorrelation ควรจะหายไป แต่หากปัญหา Autocorrelation ไม่สามารถแก้ไขได้ด้วยการเพิ่มตัวแปรอิสระที่สำคัญบางตัวเข้ามาในโมเดล อาจจะต้องพิจารณาโมเดลอื่นที่รวมเอาลักษณะความสัมพันธ์ระหว่าง ตัวแปรเข้าไว้ด้วย ในที่นี้จะกล่าวถึงวิธีการแปลงข้อมูลเพื่อแก้ปัญหา Autocorrelation สำหรับโมเดลถดถอยเชิงเส้นตรงอย่างง่ายเท่านั้น ซึ่งสามารถขยายต่อไปยังโมเดลถดถอยเชิงเส้นตรงแบบพหุได้

พิจารณาการแปลงค่าของตัวแปรตาม

$$Y_t' = Y_t - \rho Y_{t-1} \tag{4.39}$$

แทนค่า Y_t และ Y_{t-1} ด้วยโมเดลถดถอยที่สอดคล้องกัน จะได้ว่า

$$\begin{aligned} Y'_t &= (\beta_0 + \beta_1 X_t + \epsilon_t) - \rho(\beta_0 + \beta_1 X_{t-1} + \epsilon_{t-1}) \\ &= \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + (\epsilon_t - \rho\epsilon_{t-1}) \\ &= \beta'_0 + \beta'_1 X'_t + u_t \end{aligned} \tag{4.40}$$

เมื่อ

$$\begin{aligned} Y'_t &= Y_t - \rho Y_{t-1} \\ X'_t &= X_t - \rho X_{t-1} \\ u'_t &= \epsilon_t - \rho\epsilon_{t-1} \\ \beta'_0 &= \beta_0(1-\rho) \\ \beta'_1 &= \beta_1 \end{aligned}$$

จะเห็นได้ว่าโมเดล (4.40) ก็คือรูปแบบของโมเดลถดถอยเชิงเส้นตรงอย่างง่ายสำหรับตัวแปรที่มีการแปลงค่าแล้วนั่นเอง โดยที่ตัวรบกวนสุ่ม u_t เป็นอิสระกัน ทำให้ตัวประมาณกำลังสองน้อยที่สุดที่ได้จากโมเดลดังกล่าวมีคุณสมบัติที่เหมาะสม เนื่องจากไม่ทราบค่า Autocorrelation parameter ρ และเพื่อที่จะนำโมเดลของตัวแปรที่มีการแปลงค่าแล้วไปประยุกต์ใช้ จะต้องประมาณค่า ρ ก่อน โดยให้ r เป็นตัวประมาณพารามิเตอร์ ρ ดังนั้นจะได้ว่า

$$Y'_t = Y_t - rY_{t-1} \tag{4.41}$$

$$X'_t = X_t - rX_{t-1} \tag{4.42}$$

และได้สมการถดถอยของตัวแปรที่มีการแปลงค่าแล้วเป็น

$$\hat{Y}' = b'_0 + b'_1 X' \tag{4.43}$$

หากสมการ (4.43) สามารถกำจัดปัญหา Autocorrelation ได้ สมการ (4.43) สามารถแปลงกลับให้อยู่ในรูปของตัวแปรเดิมได้ดังนี้

$$\hat{Y} = b_0 + b_1 X$$

เมื่อ

$$b_0 = \frac{b'_0}{1-r} \tag{4.44}$$

$$b_1 = b'_1 \tag{4.45}$$

โดยที่ความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์ถดถอยสำหรับตัวแปรเดิมสามารถคำนวณได้จากตัวแปรที่มีการแปลงค่าแล้วดังนี้

$$S_{b_0} = \frac{S_{b'_0}}{1-r} \tag{4.46}$$

$$S_{b_1} = S_{b'_1} \tag{4.47}$$

วิธีประมาณค่าพารามิเตอร์ ρ ทำได้หลายวิธี ในที่นี้จะนำเสนอ 3 วิธี ดังนี้

1. วิธี Cochrane-Orcutt

วิธี Cochrane-Orcutt ประกอบด้วยกระบวนการทำซ้ำ 3 ขั้นตอน ดังนี้

1. การประมาณค่า ρ ทำได้โดยพิจารณาโมเดลของความคลาดเคลื่อนซึ่งอยู่ในกระบวนการ AR(1) ว่าเป็นโมเดลถดถอยที่ผ่านจุดกำเนิด

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

เมื่อ

ϵ_t แทน ตัวแปรตาม

ϵ_{t-1} แทน ตัวแปรอิสระ

u_t แทน ความคลาดเคลื่อนสุ่ม

ρ แทน ความชันของเส้นถดถอยที่ผ่านจุดกำเนิด

เนื่องจากไม่ทราบค่า ϵ_t และ ϵ_{t-1} ดังนั้นจะใช้ค่าความคลาดเคลื่อนของข้อมูลตัวอย่างที่ได้จากวิธีกำลังสองน้อยที่สุด คือ e_t และ e_{t-1} แทนตัวแปรตามและตัวแปรอิสระตามลำดับ และประมาณค่า ρ โดยสร้างสมการถดถอยเชิงเส้นตรงผ่านจุดกำเนิด และคำนวณค่าประมาณของความชัน ρ ได้เป็น

$$r = \frac{\sum_{i=2}^{n} e_{t-1}e_t}{\sum_{i=1}^{n} e_{t-1}^2} \tag{4.48}$$

2. สร้างโมเดลถดถอยที่มีการแปลงค่าแล้ว โดยแทนค่า r ลงในสมการ (4.41) และ (4.42) แล้วสร้างสมการถดถอยของตัวแปรที่มีการแปลงค่าแล้ว Y' และ X' ด้วยวิธีกำลังสองน้อยที่สุด

3. ใช้สถิติ Durbin-Watson ทดสอบว่าความคลาดเคลื่อนของสมการที่มีการแปลงค่าแล้วมีความสัมพันธ์กันหรือไม่ ถ้าผลการทดสอบชี้ว่าความคลาดเคลื่อนเป็นอิสระกัน ก็จะยอมรับสมการที่มีการแปลงค่าแล้ว จากนั้นแปลงค่าสัมประสิทธิ์ถดถอยกลับให้อยู่ในรูปของตัวแปรเดิม แต่ถ้าผลการทดสอบชี้ว่าปัญหา Autocorrelation ยังคงมีอยู่ ก็จะทำต่อในรอบที่สอง โดยประมาณค่า ρ ใหม่ ใช้สมการถดถอยที่แปลงกลับให้อยู่ในรูปของตัวแปรเดิมที่ได้จากข้อ 2 จากนั้นสร้างสมการถดถอยใหม่ และตรวจสอบความเป็นอิสระกันของความคลาดเคลื่อนอีกครั้ง หากปัญหา Autocorrelation ยังคงมีอยู่ ก็จะทำรอบที่ $3, 4, \ldots$ ต่อไปเรื่อย ๆ จนกระทั่งแก้ปัญหา Autocorrelation ได้ แต่อย่างไรก็ตามวิธีนี้อาจไม่สามารถแก้ไขปัญหา Autocorrelation ได้เสมอไป ซึ่งกรณีเช่นนี้อาจต้องพิจารณาวิธีอื่นแทน

หมายเหตุ สถิติ Durbin-Watson มีความสัมพันธ์กับค่าประมาณพารามิเตอร์ ρ ดังนี้

จาก

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$$

เนื่องจาก $\sum_{t=2}^{n} e_t^2 \approx \sum_{t=2}^{n} e_{t-1}^2$ จะได้ว่า

$$d \approx 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}\right)$$

จากสมการ (4.48) ได้ว่า $r = \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$

ดังนั้น

$$d \approx 2(1 - r)$$

เนื่องจาก $-1 \leq r \leq 1$ ดังนั้นสถิติ Durbin-Watson จะมีค่าอยู่ระหว่าง 0 ถึง 4 โดยประมาณ และ $d \approx 2$ เมื่อ $r = 0$

ตัวอย่างที่ 4.10 จากข้อมูลในตัวอย่างที่ 4.9 จงแก้ปัญหา Positive autocorrelation โดยใช้วิธี Cochrane-Orcutt

ปีพ.ศ.	ไตรมาสที่	Y_t	X_t	e_t	e_{t-1}	e_te_{t-1}	e^2_{t-1}
2536	1	20.96	127.3	-0.0261	-	-	-
	2	21.40	130.0	-0.0620	-0.0261	0.00162	0.00068
	3	21.96	132.7	0.0220	-0.0620	-0.00137	0.00385
	4	21.52	129.4	0.1638	0.0220	0.00361	0.00048
2537	1	22.39	135.0	0.0466	0.1638	0.00763	0.02682
	2	22.76	137.1	0.0464	0.0466	0.00216	0.00217
	3	23.48	141.2	0.0436	0.0464	0.00202	0.00215
	4	23.66	142.8	-0.0584	0.0436	-0.00255	0.00190
2538	1	24.10	145.5	-0.0944	-0.0584	0.00552	0.00341
	2	24.01	145.3	-0.1491	-0.0944	0.01408	0.00891
	3	24.54	148.3	-0.1480	-0.1491	0.02207	0.02224
	4	24.30	146.4	-0.0531	-0.1480	0.00785	0.02190
2539	1	25.00	150.2	-0.0229	-0.0531	0.00122	0.00281
	2	25.64	153.1	0.1059	-0.0229	-0.00243	0.00053
	3	26.36	157.3	0.0855	0.1059	0.00905	0.01120
	4	26.98	160.7	0.1061	0.0855	0.00907	0.00730
2540	1	27.52	164.2	0.0291	0.1061	0.00309	0.01126
	2	27.78	165.6	0.0423	0.0291	0.00123	0.00085
	3	28.24	168.7	-0.0442	0.0423	-0.00187	0.00179
	4	28.78	171.7	-0.0330	-0.0442	0.00146	0.00195
	รวม	491.38	2952.5	0.0000	0.03301	0.08345	0.13221

ตารางที่ 4.16: ข้อมูลยอดขาย (Y_t) และค่าใช้จ่ายในการโฆษณา (X_t) แสดงวิธีแก้ไขปัญหา Autocorrelation

วิธีทำ จากตาราง 4.16 ประมาณค่า ρ ได้ดังนี้

$$\begin{aligned} r &= \frac{\sum_{i=2}^{n} e_{t-1}e_t}{\sum_{i=1}^{n} e^2_{t-1}} \\ &= \frac{0.08345}{0.13221} \\ &= 0.63119 \end{aligned}$$

ทำการแปลงข้อมูลได้ดังนี้

$$\begin{aligned} Y'_t &= Y_t - 0.63119Y_{t-1} \\ X'_t &= X_t - 0.63119X_{t-1} \end{aligned}$$

ปีพ.ศ.	ไตรมาสที่	Y_t	X_t	Y'_t	X'_t
2536	1	20.96	127.3	-	-
	2	21.40	130.0	8.1703	49.6495
	3	21.96	132.7	8.4525	50.6453
	4	21.52	129.4	7.6591	45.6411
2537	1	22.39	135.0	8.8068	53.3240
	2	22.76	137.1	8.6277	51.8894
	3	23.48	141.2	9.1141	54.6639
	4	23.66	142.8	8.8397	53.6760
2538	1	24.10	145.5	9.1660	55.3661
	2	24.01	145.3	8.7983	53.4619
	3	24.54	148.3	9.3851	56.5881
	4	24.30	146.4	8.8106	52.7945
2539	1	25.00	150.2	9.6621	57.7938
	2	25.64	153.1	9.8603	58.2953
	3	26.36	157.3	10.1763	60.6648
	4	26.98	160.7	10.3418	61.4138
2540	1	27.52	164.2	10.4905	62.7678
	2	27.78	165.6	10.4097	61.9586
	3	28.24	168.7	10.7055	64.1749
	4	28.78	171.7	10.9552	65.2182
	รวม	491.38	2952.5	178.4315	1069.9868

ตารางที่ 4.17: ข้อมูลยอดขาย (Y_t) และค่าใช้จ่ายในการโฆษณา (X_t) แสดงวิธีแปลงค่าตัวแปร

จากนั้นสร้างสมการถดถอยของตัวแปรที่แปลงค่าแล้ว ดังแสดงในตาราง 4.17 ได้สมการเป็น

$$\hat{Y}'_t = \underset{(0.1673)}{-0.3940} + \underset{(0.00296)}{0.1738}\, X'_t$$

ทำการตรวจสอบ Positive autocorrelation โดยกำหนดสมมติฐานดังนี้

$$H_0 : \rho = 0 \qquad vs. \qquad H_1 : \rho > 0$$

คำนวณสถิติทดสอบ Durbin-Watson ได้เป็น $d = 1.65$

จากตารางค่าวิกฤติของ Durbin-Watson ที่ $\alpha = 0.05, k = 1$ และ $n = 19$ ได้ค่า $d_L = 1.18$ และ $d_U = 1.40$ เนื่องจาก $d = 1.65 > d_U$ จึงไม่สามารถปฏิเสธ H_0 ได้ นั่นคือ ความคลาดเคลื่อนเป็นอิสระกัน เมื่อขจัดปัญหาของ autocorrelation ได้แล้ว ให้ทำการแปลงตัวแปรกลับ เพื่อสร้างสมการถดถอยของตัวแปรเดิม

$$b_1 = b'_1 = 0.1738, \qquad S_{b_1} = S_{b'_1} = 0.00296$$

$$b_0 = \frac{b'_0}{1-r} = \frac{-0.3940}{1-0.63119} = -1.0683, \qquad S_{b_0} = \frac{S_{b'_0}}{1-r} = \frac{0.1673}{1-0.63119} = 0.4536$$

และได้สมการถดถอยในรูปของตัวแปรเดิมดังนี้

$$\hat{Y}_t = -1.0683 + 0.1738X_t$$

2. วิธี Hildreth-Lu

การประมาณพารามิเตอร์ ρ ด้วยวิธี Hildreth-Lu จะคล้ายคลึงกับการแปลงข้อมูลด้วยวิธีของ Box-Cox เพื่อหา λ ที่เหมาะสมต่อการแปลงค่าของตัวแปรตาม Y ใน Power transformation โดยวิธี Hildreth-Lu จะใช้หลักการค้นหาตัวประมาณของ ρ เชิงตัวเลข โดยสมมติค่าของ ρ ขึ้นมาชุดหนึ่ง ในแต่ละค่า ρ จะทำการแปลงค่าของตัวแปรตามและตัวแปรอิสระ แล้วสร้างสมการถดถอยสำหรับตัวแปรที่แปลงค่าแล้ว พร้อมกับคำนวณค่า SSE โดยตัวประมาณของพารามิเตอร์ ρ ก็คือ ค่าที่ทำให้ SSE มีค่าต่ำที่สุดนั่นเอง

เมื่อได้ค่าประมาณของพารามิเตอร์ ρ แล้ว สร้างสมการถดถอยที่สอดคล้องกับค่าประมาณดังกล่าว จากนั้นใช้สถิติ Durbin-Watson ตรวจสอบว่าความคลาดเคลื่อนของสมการที่สร้างขึ้นเป็นอิสระกันหรือไม่ หากสมการที่ได้สามารถขจัดปัญหา Autocorrelation ได้แล้ว สร้างสมการถดถอยให้อยู่ในรูปของตัวแปรเดิม โดยแทนค่าเข้าในสมการ (4.44) และ (4.45) ตามลำดับ

หมายเหตุ วิธี Hildreth-Lu แตกต่างจากวิธี Cochrane-Orcutt ในแง่ที่วิธี Hildreth-Lu ไม่ต้องมีการทำซ้ำ เมื่อที่ได้ค่าประมาณของพารามิเตอร์ ρ แล้ว

3. วิธี First-Difference

จากสมการ (4.40)

$$Y'_t = \beta'_0 + \beta'_1 X'_t + u_t$$

เมื่อ $\beta'_0 = \beta_0(1-\rho)$ และ $\beta'_1 = \beta_1$

หาก $\rho = 1$ จะได้ว่า $\beta'_0 = \beta_0(1-\rho) = 0$ และโมเดลที่แปลงค่าแล้วจะลดรูปเป็น

$$Y'_t = \beta'_1 X'_t + u_t \qquad (4.49)$$

เมื่อ

$$Y_t' = Y_t - Y_{t-1} \quad (4.50)$$

$$X_t' = X_t - X_{t-1} \quad (4.51)$$

$$\beta_1' = \beta_1 \quad (4.52)$$

จะเห็นได้ว่าตัวแปรที่แปลงค่าแล้วจะอยู่ในรูปของผลต่างลำดับที่หนึ่ง หรือ First difference นั่นเอง จากนั้นทำการประมาณค่าพารามิเตอร์ β_1' ในสมการถดถอยผ่านจุดกำเนิดด้วยวิธีกำลังสองน้อยที่สุด ได้สมการสำหรับตัวแปรที่แปลงค่าด้วยวิธี First difference ดังนี้

$$\hat{Y}_t' = b_1' X_t' \quad (4.53)$$

โดยที่สมการ (4.53) สามารถแปลงกลับให้อยู่ในรูปของตัวแปรเดิมได้เป็น

$$\hat{Y}_t = b_0 + b_1 X_t$$

เมื่อ

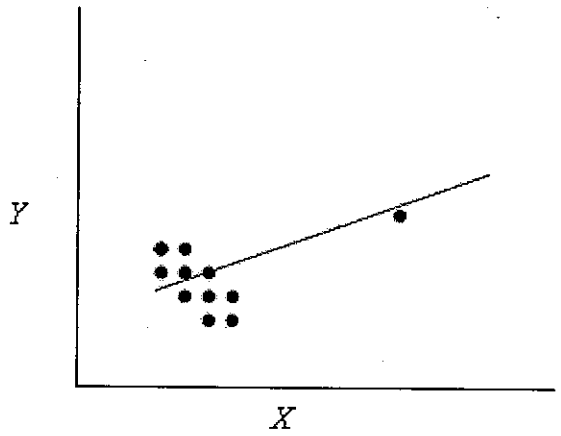
$$b_0 = \bar{Y} - b_1' \bar{X} \quad (4.54)$$

$$b_1 = b_1' \quad (4.55)$$

หมายเหตุ วิธีแปลงข้อมูลทั้งสามวิธีอาจให้ผลที่ไม่สอดคล้องกัน แต่หากวิธีแปลงข้อมูลทั้ง 3 วิธีสามารถขจัดปัญหา Autocorrelation ได้ อาจพิจารณาเลือกวิธีที่ง่ายต่อการคำนวณ

4.7 การตรวจสอบค่าสังเกตที่มีอิทธิพล (Influential Observations)

บ่อยครั้งที่ข้อมูลกลุ่มเล็ก ๆ อาจมีอิทธิพลต่อโมเดลถดถอยอย่างไม่เป็นสัดส่วนกับข้อมูลส่วนใหญ่ นั่นคือ ค่าประมาณพารามิเตอร์หรือค่าพยากรณ์อาจจะขึ้นอยู่กับกลุ่มข้อมูลที่มีอิทธิพลนั้นมากกว่าข้อมูลส่วนใหญ่ รูปที่ 4.17 แสดงถึงค่าสังเกตที่มีอิทธิพล ซึ่งจะเห็นได้ว่าค่าดังกล่าวมีอิทธิพลต่อรูปแบบของสมการถดถอยมากกว่าข้อมูลส่วนใหญ่ หากตัดค่าดังกล่าวออกไป ความชันของสมการถดถอยจะเปลี่ยนจากบวกเป็นลบ ดังนั้นการค้นหาค่าสังเกตที่มีอิทธิพลจึงมีความสำคัญ เพราะจะทำให้สามารถตรวจสอบอิทธิพลของค่าดังกล่าวต่อโมเดลได้ ถ้าค่าสังเกตที่มีอิทธิพลนั้นเกิดจากความผิดพลาดในการเก็บข้อมูลหรือจดบันทึกข้อมูล อาจตัดค่าดังกล่าวออกไปหรือแก้ไขให้ถูกต้องก่อนที่จะสร้างสมการถดถอยใหม่ ในทางตรงกันข้ามหากไม่พบสิ่งผิดปกติของค่าสังเกตที่



รูปที่ 4.17: กราฟแสดงค่าสังเกตที่มีอิทธิพล

มีอิทธิพลนั้น แต่พบว่าค่าดังกล่าวมีอิทธิพลต่อโมเดลอย่างมาก ก็ควรที่จะค้นหาสาเหตุว่าเนื่องจากเหตุใด ซึ่งวิธีตรวจสอบค่าสังเกตที่มีอิทธิพลมีหลายวิธี โดยจะกล่าวถึงรายละเอียดในหัวข้อต่อไปนี้

4.7.1 ค่า Leverage หรือ Hat Matrix

ตำแหน่งของค่าสังเกตในสเปซของ X มีความสำคัญต่อการกำหนดคุณสมบัติของโมเดล โดยเฉพาะค่าสังเกตที่ตกห่างจากค่าอื่นอาจมีน้ำหนักที่ไม่ได้สัดส่วน ซึ่งอาจมีผลกระทบต่อการประมาณค่าพารามิเตอร์ การพยากรณ์ และการคำนวณสถิติพื้นฐาน Hoaglin และ Welsch (1978) ได้ศึกษาการใช้ Hat matrix ซึ่งมีรูปแบบทั่วไปเป็น $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ในการค้นหาค่าสังเกตที่มีอิทธิพล โดยสมาชิก h_{ij} ของ Hat matrix แสดงถึงน้ำหนักของ Y_j บน $\hat{Y}_i$ ซึ่งการตรวจสอบสมาชิกของ Hat matrix อาจชี้ถึงค่าสังเกตที่มีอิทธิพลได้ โดยจะพิจารณาจากตำแหน่งของค่าดังกล่าวในสเปซของ X โดยเฉพาะค่าที่อยู่ในแนวทแยงมุมของ Hat matrix หรือ h_{ii} และเรียก h_{ii} ว่า *Leverage* ของค่าสังเกตที่ i เนื่องจาก

$$\hat{\mathbf{Y}} = \mathbf{HY}, \qquad V(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$$
$$\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{Y}, \qquad V(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$$

จะได้ว่า $\hat{Y}_i$ เป็นผลรวมเชิงเส้น (Linear combination) ของ Y_i และ h_{ii} ก็คือน้ำหนักของค่าสังเกต Y_i ที่มีต่อค่าที่อยู่บนเส้นถดถอย $\hat{Y}_i$ หาก h_{ii} มีค่ามาก แสดงว่า Y_i มีอิทธิพลต่อการกำหนด $\hat{Y}_i$ มาก เนื่องจาก h_{ii} เป็นฟังก์ชันของ X ดังนั้นจะใช้วัดบทบาทของ X ในการกำหนดความสำคัญของ Y_i ที่มีอิทธิพลต่อ $\hat{Y}_i$ นอกจากนี้เมื่อ h_{ii} มีค่ามาก ยังทำให้ความแปรปรวนของความคลาดเคลื่อน e_i มีค่าลดลงด้วย แสดงว่าค่าพยากรณ์มีค่าใกล้เคียงกับค่าที่แท้จริงมากขึ้น หาก $h_{ii} = 1$ แล้ว ทำให้ $V(e_i) = 0$ นั่นคือ $\hat{Y}_i = Y_i$ โดย h_{ii} มีคุณสมบัติที่สำคัญดังนี้

1. $0 \leq h_{ii} \leq 1$

2. $\sum_{i=1}^{n} h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p$
   เมื่อ p แทน จำนวนพารามิเตอร์ในสมการถดถอยที่มีระยะตัดแกน Y

3. ค่าเฉลี่ยของสมาชิกที่อยู่ในแนวทแยงมุมของ Hat matrix มีค่าเป็น

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

กรณีสมการถดถอยเชิงเส้นตรงอย่างง่าย จะได้สมการของ Hat matrix เป็น

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4.56}$$

จะเห็นได้ว่า h_{ii} เป็นค่าที่ใช้วัดระยะทางของค่าสังเกตตัวที่ i โดยจะพิจารณาจากผลต่างระหว่างค่า X กับค่าเฉลี่ยของมัน หาก h_{ii} มีค่ามาก แสดงว่าค่าสังเกตที่ i อยู่ห่างจากจุดศูนย์กลางของ X มาก

โดยทั่วไปจะถือว่าค่าสังเกตที่มี $h_{ii} > \frac{2p}{n}$ หรือ $\frac{3p}{n}$ เป็นค่าสังเกตที่มีอิทธิพล (High leverage points) นอกจากนี้ยังมีเกณฑ์ในการพิจารณาอื่นซึ่งระบุว่าถ้า $h_{ii} > 0.5$ จัดเป็นค่าสังเกตที่มีอิทธิพลมาก และ $0.2 < h_{ii} < 0.5$ จัดเป็นค่าสังเกตที่มีอิทธิพลปานกลาง

4.7.2 Cook's Distance

จะเห็นได้ว่าค่า Leverage สามารถใช้ค้นหาค่าสังเกตที่มีอิทธิพลได้โดยคำนึงถึงตำแหน่งของค่าสังเกตนั้นในสเปซของ X เพียงอย่างเดียว ดังนั้นวิธีที่จะใช้ค้นหาค่าสังเกตที่มีอิทธิพลที่พิจารณาทั้งตำแหน่งของค่าสังเกตและตัวแปรตามจึงเป็นที่ต้องการอย่างมาก Cook (1977, 1978) ได้เสนอให้วัดอิทธิพลของค่าสังเกตโดยใช้ระยะทางกำลังสอง (Squared distance) ของค่าสังเกตที่ i ที่มีต่อค่าพยากรณ์ทั้ง n ค่ากับค่าพยากรณ์ที่ได้เมื่อไม่รวมค่าสังเกตที่ i ที่สอดคล้องกัน ให้ D_i แทน Cook's distance ซึ่งใช้วัดอิทธิพลของค่าสังเกตโดยรวม มีสูตรดังนี้

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pS^2} \tag{4.57}$$

เมื่อ $\hat{Y}_{j(i)}$ แทน ค่าพยากรณ์ที่ได้จากสมการถดถอยเมื่อตัดค่าสังเกตที่ i ออกไป

สมการ (4.57) สามารถเขียนให้อยู่ในรูปของเมตริกซ์ได้ดังนี้

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{pS^2} \tag{4.58}$$

เมื่อ

$\hat{\mathbf{Y}}$ แทน เวกเตอร์ของค่าพยากรณ์ที่ได้จากสมการถดถอยประกอบด้วยค่าสังเกต n ค่า

$\hat{\mathbf{Y}}_{(i)}$ แทน เวกเตอร์ของค่าพยากรณ์ที่ได้จากสมการถดถอยเมื่อตัดค่าสังเกตที่ i ออกไป

นอกจากนี้สมการ (4.58) ยังสามารถเขียนให้อยู่ในรูปของระยะทางกำลังสองระหว่างตัวประมาณกำลังสองน้อยที่สุดได้ดังนี้

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(i)})}{pS^2} \tag{4.59}$$

เมื่อ

$\mathbf{b}$ แทน เวกเตอร์ค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากค่าสังเกต n ค่า

$\mathbf{b}_{(i)}$ แทน เวกเตอร์ค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากการตัดค่าสังเกตที่ i ออกไป

ค่าสังเกตที่มีค่า D_i มาก ๆ แสดงว่าเป็นค่าสังเกตที่มีอิทธิพล โดยอาจเปรียบเทียบขนาดของ D_i กับ $F_{p,n-p\,(\alpha)}$ ถ้า $D_i \approx F_{p,n-p\,(.5)}$ แล้ว การตัดค่าสังเกตที่ i จะทำให้ $\mathbf{b}$ มีค่าเข้าใกล้ขอบเขตของบริเวณเชื่อมั่น (Confidence region) ขนาด 50% ของ β ที่ได้จากข้อมูลทั้ง n ค่า ดังนั้นการเปลี่ยนตำแหน่งของ $\mathbf{b}$ อย่างมาก จะแสดงถึงอิทธิพลของค่าสังเกตที่ i ที่มีต่อตัวประมาณกำลังสองน้อยที่สุด เนื่องจาก $F_{p,n-p\,(.5)} \approx 1$ ดังนั้นจะถือว่าค่าสังเกตที่มี $D_i > 1$ เป็นค่าสังเกตที่มีอิทธิพล โดยทั่วไปค่าสังเกตที่ตกอยู่ในขอบเขตของบริเวณเชื่อมั่นขนาด $10-20\%$ จะจัดเป็นค่าสังเกตที่มีอิทธิพลต่อค่าพยากรณ์ค่อนข้างน้อย และค่าสังเกตที่ตกอยู่ในขอบเขตของบริเวณเชื่อมั่นขนาด 50% หรือมากกว่า จะจัดเป็นค่าสังเกตที่มีอิทธิพลต่อค่าพยากรณ์ค่อนข้างมาก

นอกจากนี้ Cook's distance ยังสามารถคำนวณโดยไม่ต้องสร้างสมการถดถอยใหม่ทุกครั้งที่ค่าสังเกตที่ i ถูกตัดออกไป โดยมีรูปแบบสมการที่เทียบเท่ากันดังนี้

$$D_i = \frac{e_i^2}{pS^2}\left\{\frac{h_{ii}}{(1-h_{ii})^2}\right\} \tag{4.60}$$

จะเห็นได้ว่า D_i ในสมการ (4.60) ขึ้นอยู่กับขนาดของความคลาดเคลื่อน e_i และค่า Leverage h_{ii} ซึ่งใช้วัดความเหมาะสมของโมเดลในการพยากรณ์ค่าสังเกตที่ i และวัดระยะห่างระหว่างค่าสังเกตดังกล่าวกับจุดศูนย์-

กลางของข้อมูลที่เหลือ ถ้า e_i หรือ h_{ii} หรือทั้งสองตัวมีค่ามาก จะทำให้ D_i มีค่ามากตามไปด้วย และส่งผลให้ค่าสังเกตที่ i เป็นค่าสังเกตที่มีอิทธิพล

4.7.3 *DFFITS*

ใช้วัดอิทธิพลของค่าสังเกตที่ i ที่มีต่อค่าพยากรณ์ $\hat{Y}_i$ มีสูตรทั่วไปดังนี้

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}, \qquad i = 1, 2, \ldots, n \tag{4.61}$$

เมื่อ

$\hat{Y}_i$ แทน ค่าพยากรณ์ที่ได้จากสมการถดถอยประกอบด้วยค่าสังเกต n ค่า

$\hat{Y}_{(i)}$ แทน ค่าพยากรณ์ที่ได้จากสมการถดถอยเมื่อตัดค่าสังเกตที่ i ออกไป

$S^2_{(i)}$ แทน ส่วนเบี่ยงเบนมาตรฐานของสมการถดถอยเมื่อตัดค่าสังเกตที่ i ออกไป โดยที่

$$S^2_{(i)} = \frac{(n-p)S^2 - e_i^2/(1-h_{ii})}{n-p-1} \tag{4.62}$$

จะเห็นได้ว่า *DFFITS* เป็นการหาผลต่างของค่าพยากรณ์ที่ได้จากสมการถดถอยที่มีค่าสังเกต n ค่า และจากสมการถดถอยที่ตัดค่าสังเกตที่ i ออกไป ส่วนตัวหารในสมการ (4.61) แทน ค่าประมาณของส่วนเบี่ยงเบนมาตรฐานของ $\hat{Y}_i$ แต่ใช้ $S^2_{(i)}$ ที่ได้จากการตัดค่าสังเกตที่ i ออกจากการประมาณความแปรปรวนของความคลาดเคลื่อน σ^2 ดังนั้น *DFFITS* จึงเป็นค่าที่ใช้วัดการเปลี่ยนแปลงของค่าพยากรณ์เมื่อค่าสังเกตที่ i ถูกตัดออกไปนั่นเอง

ทำนองเดียวกัน *DFFITS* ยังสามารถคำนวณได้จากค่าสังเกตทั้ง n ค่า โดยไม่ต้องสร้างสมการใหม่ทุกครั้งที่ตัดค่าสังเกตออกไป มีสูตรดังนี้

$$DFFITS_i = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \tag{4.63}$$

เมื่อ $t_i = \dfrac{e_i}{S_{(i)}\sqrt{1-h_{ii}}}$ แทน ค่า *Externally studentized residual* และนิยมเรียกย่อ ๆ ว่า *R-student*

โดย t_i จะแตกต่างจาก Studentized residual r_i หากค่าสังเกตที่ i มีอิทธิพลแล้ว $S^2_{(i)}$ จะมีค่าแตกต่างจาก S^2 มาก โดยถ้าข้อมูลมีลักษณะผิดปกติ จะทำให้ t_i มีค่ามาก ขณะเดียวกันถ้าข้อมูลมีค่า Leverage สูง ๆ

นั่นคือ h_{ii} มีค่าเข้าใกล้ 1 จะทำให้ $DFFITS_i$ มีค่ามากตามไปด้วย แต่อย่างไรก็ตามถ้า $h_{ii} \approx 0$ พบว่า t_i จะมีอิทธิพลต่อ $DFFITS_i$ ปานกลาง ทำนองเดียวกันถ้า t_i มีค่าเข้าใกล้ 0 แต่ h_{ii} มีค่ามาก ยังคงส่งผลกระทบต่อ $DFFITS_i$ ปานกลาง ดังนั้น $DFFITS_i$ จะมีค่ามาก ต้องถูกกระทบด้วยค่า Leverage และ R-student พร้อม ๆ กัน

โดยทั่วไปจะถือว่าค่าสังเกตที่ i เป็นค่าสังเกตที่มีอิทธิพล ถ้า $|DFFITS_i| > 1$ สำหรับข้อมูลขนาดเล็กถึงปานกลาง และ $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ สำหรับข้อมูลขนาดใหญ่

4.7.4 *DFBETAS*

เป็นค่าที่ใช้วัดอิทธิพลของค่าสังเกตที่ i ที่มีต่อค่าสัมประสิทธิ์ถดถอยแต่ละตัว b_j เมื่อ $j = 0, 1, 2, \ldots, k$ โดยหาผลต่างระหว่างค่าประมาณของสัมประสิทธิ์ถดถอย b_j ที่ได้จากค่าสังเกต n ค่า กับค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากการตัดค่าสังเกตที่ i ออกไป นั่นคือ $DFBETAS_{j(i)}$ จะวัดการเปลี่ยนแปลงของ b_j ในหน่วยของส่วนเบี่ยงเบนมาตรฐาน เมื่อค่าสังเกตที่ i ถูกตัดออกไป มีสูตรทั่วไปดังนี้

$$DFBETAS_{j(i)} = \frac{b_j - b_{j(i)}}{S^2_{(i)} C_{jj}} \tag{4.64}$$

เมื่อ

c_{jj} แทน ค่าที่ j ที่อยู่ในแนวทแยงมุมของเมตริกซ์ $(\mathbf{X}'\mathbf{X})$

$b_{j(i)}$ แทน ค่าสัมประสิทธิ์ถดถอยตัวที่ j ที่ได้จากการตัดค่าสังเกตที่ i ออกไป

ทำนองเดียวกัน $DFBETAS_{j(i)}$ สามารถคำนวณได้โดยที่ไม่ต้องสร้างสมการถดถอยใหม่ทุกครั้งที่ตัดค่าสังเกตออกไป มีสูตรดังนี้

$$\begin{aligned} DFBETAS_{j(i)} &= \frac{r_{ji}}{\sqrt{\mathbf{r}_j'\mathbf{r}_j}} \cdot \frac{e_i}{S_{(i)}(1-h_{ii})} \\ &= \frac{r_{ji}}{\sqrt{\mathbf{r}_j'\mathbf{r}_j}} \cdot \frac{t_i}{\sqrt{1-h_{ii}}} \end{aligned} \tag{4.65}$$

เมื่อ

t_i แทน ค่า R-student

r_{ji} แทน สมาชิกที่อยู่ในตำแหน่งที่ (j, i) ของเมตริกซ์ $\underset{p\times n}{\mathbf{R}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$\mathbf{r}_j$ แทน สมาชิกในแถวที่ j ของเมตริกซ์ $\mathbf{R}$ ซึ่งก็คือ Leverage ของค่าสังเกต n ค่าที่มีต่อ b_j นั่นเอง

โดยเครื่องหมายของ $DFBETAS_{j(i)}$ ชี้ว่าการรวมค่าสังเกตที่ i ไว้ในการวิเคราะห์ จะทำให้ค่าประมาณของสัมประสิทธิ์ถดถอยที่ j เพิ่มขึ้นหรือลดลง และขนาดของ $DFBETAS_{j(i)}$ ในค่าสัมบูรณ์หมายถึงขนาดของความแตกต่างระหว่างค่าสัมประสิทธิ์ถดถอย เมื่อเทียบกับค่าประมาณส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์ถดถอย ดังนั้นถ้า $|DFBETAS_{j(i)}|$ มีค่ามาก แสดงว่าค่าสังเกตที่ i มีอิทธิพลต่อสัมประสิทธิ์ถดถอยตัวที่ j โดยทั่วไปจะถือว่าค่าสังเกตที่ i เป็นค่าสังเกตที่มีอิทธิพล ถ้า $|DFBETAS_{j(i)}| > 1$ สำหรับข้อมูลขนาดเล็กถึงปานกลาง และ $|DFBETAS_{j(i)}| > \frac{2}{\sqrt{n}}$ สำหรับข้อมูลขนาดใหญ่

4.7.5 *COVRATIOS*

การพิจารณาค่า Cook's distance, $DFFITS$ และ $DFBETAS$ ที่ได้กล่าวมาแล้ว เป็นการตรวจสอบอิทธิพลของค่าสังเกตแต่ละค่าที่มีต่อค่าประมาณของสัมประสิทธิ์ถดถอย b_j และค่าพยากรณ์ $\hat{Y}_i$ เท่านั้น แต่ไม่ได้ให้ข้อมูลเกี่ยวกับความแม่นยำของการประมาณ ปกติแล้วการวัดความแม่นยำของการประมาณ นิยมคำนวณจากค่าดีเทอร์มิแนนท์ของเมตริกซ์ความแปรปรวนร่วม (Covariance matrix) ซึ่งเรียกว่า *Generalized variance* ดังนั้นการตรวจสอบอิทธิพลของค่าสังเกตที่ i ที่มีต่อความแม่นยำของการประมาณ จะคำนวณจาก

$$COVRATIO_i = \frac{\left|\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1} S^2_{(i)}\right|}{\left|(\mathbf{X}'\mathbf{X})^{-1} S^2\right|}, \qquad i = 1, 2, \ldots, n \tag{4.66}$$

เมื่อ $\mathbf{X}_{(i)}$ แทน เมตริกซ์ของตัวแปรอิสระที่เกิดจากการตัดแถวที่ i ออกไปแล้ว

จะเห็นได้ว่าถ้า $COVRATIO_i > 1$ แสดงว่าค่าสังเกตที่ i ช่วยเพิ่มความแม่นยำในการประมาณในทางตรงข้ามถ้า $COVRATIO_i < 1$ แสดงว่าการรวมค่าสังเกตที่ i จะทำให้ความแม่นยำในการประมาณลดลง ทำนองเดียวกัน $COVRATIO_i$ สามารถคำนวณได้จากสูตรที่เทียบเท่ากันดังนี้

$$COVRATIO_i = \frac{\left(S^2_{(i)}\right)^p}{S^p} \cdot \left(\frac{1}{1 - h_{ii}}\right) \tag{4.67}$$

จากสมการที่ (4.67) พบว่าค่าสังเกตที่มี Leverage มาก จะทำให้ $COVRATIO_i$ มีค่ามากตามไปด้วย เนื่องจากค่าสังเกตที่มี Leverage มาก จะช่วยเพิ่มความแม่นยำของการประมาณ หากค่าสังเกตดังกล่าวไม่เป็นค่าสังเกตที่ผิดปกติในสเปซของ Y และถ้าค่าสังเกตที่ i เป็นค่าสังเกตที่ผิดปกติ จะทำให้ $\frac{\left(S^2_{(i)}\right)^p}{S^p}$ มีค่าน้อยกว่า 1

โดยทั่วไปจะถือว่าค่าสังเกตที่ i เป็นค่าสังเกตที่มีอิทธิพล ถ้า $|COVRATIO_i| > 1 + \frac{3p}{n}$ หรือ $|COVRATIO_i| < 1 - \frac{3p}{n}$ โดยที่ขอบเขตล่างของเกณฑ์การพิจารณาจะเหมาะสมก็ต่อเมื่อ $n > 3p$ และ

เกณฑ์ดังกล่าวเหมาะสำหรับข้อมูลขนาดใหญ่

ตัวอย่างที่ 4.11 ข้อมูลต่อไปนี้แสดงเวลาที่ใช้ในการขนส่งสารเคมีจำนวน 25 ครั้ง (Y) หน่วยเป็นนาที จำนวนกล่องสารเคมี (X_1) และน้ำหนักสารเคมีทั้งหมด (X_2) หน่วยเป็นปอนด์ จงตรวจสอบว่าข้อมูลชุดนี้มีค่าสังเกตที่ผิดปกติและค่าสังเกตที่มีอิทธิพลหรือไม่

ครั้งที่	Y	X_1	X_2
1	15.53	8	535
2	10.35	4	195
3	10.88	4	315
4	13.73	5	55
5	12.60	7	125
6	16.96	8	305
7	6.85	3	85
8	16.68	8	185
9	78.09	31	1435
10	20.35	6	580
11	39.18	17	663
12	19.85	11	190
13	12.35	5	230
14	18.60	7	437
15	22.85	10	423
16	27.85	11	751
17	14.20	7	175
18	17.85	8	107
19	8.35	4	11
20	33.95	18	745
21	16.75	11	115
22	51.17	27	785
23	17.60	10	425
24	18.68	9	610
25	9.60	5	125

ตารางที่ 4.18: ข้อมูลสำหรับตัวอย่างที่ 4.11

วิธีทำ จากข้อมูลในตาราง 4.19 คำนวณค่า h_{ii}, D_i, $DFFITS$, $DFBETAS$ และ $COVRATIO$ ได้ดังแสดงในตาราง 4.19 แล้วจึงคำนวณเกณฑ์ที่ใช้ในการพิจารณาค่าสังเกตที่ผิดปกติและค่าสังเกตที่มีอิทธิพลดังนี้

ครั้งที่	Y	$\hat{Y}$	e_i	D_i	t_i	h_{ii}	COV $RATIO$	$DFFITS$	$DFBETAS$		
									Intercept	X_1	X_2
1	15.53	20.5581	-5.0281	0.100	-1.6956	0.1018	0.8711	-0.5709	-0.2745	0.4113	-0.4349
2	10.35	9.2036	1.1464	0.003	0.3575	0.0707	1.2149	0.0986	0.0941	-0.0478	0.0144
3	10.88	10.9298	-0.0498	0.000	-0.0157	0.0987	1.2757	-0.0052	-0.0042	0.0039	-0.0028
4	13.73	8.8056	4.9244	0.078	1.6392	0.0854	0.8760	0.5008	0.3975	0.0883	-0.2734
5	12.6	13.0444	-0.4444	0.001	-0.1386	0.0750	1.2396	-0.0395	-0.0264	-0.0133	0.0242
6	16.96	17.2496	-0.2896	0.000	-0.0887	0.0429	1.1999	-0.0188	-0.0142	0.0018	0.0011
7	6.85	6.0054	0.8446	0.002	0.2646	0.0818	1.2398	0.0790	0.0771	-0.0223	-0.0110
8	16.68	15.5234	1.1566	0.003	0.3594	0.0637	1.2056	0.0938	0.0589	0.0334	-0.0538
9	78.09	70.6703	7.4197	3.419	4.3108	0.4983	0.3422	4.2961	-2.4824	0.9287	1.5076
10	20.35	17.9736	2.3764	0.054	0.8068	0.1963	1.3054	0.3987	0.1804	-0.3382	0.3413
11	39.18	36.9425	2.2375	0.016	0.7099	0.0861	1.1717	0.2180	-0.0468	0.0925	-0.0027
12	19.85	20.4430	-0.5930	0.002	-0.1890	0.1137	1.2906	-0.0677	-0.0174	-0.0487	0.0540
13	12.35	11.3230	1.0270	0.002	0.3185	0.0611	1.2070	0.0813	0.0754	-0.0356	0.0113
14	18.6	17.5325	1.0675	0.003	0.3342	0.0782	1.2277	0.0974	0.0622	-0.0671	0.0618
15	22.85	22.1788	0.6712	0.001	0.2057	0.0411	1.1918	0.0426	0.0226	-0.0048	0.0068
16	27.85	28.5129	-0.6629	0.003	-0.2178	0.1659	1.3692	-0.0972	-0.0188	0.0644	-0.0842
17	14.2	13.7636	0.4364	0.000	0.1349	0.0594	1.2192	0.0339	0.0254	0.0065	-0.0157
18	17.85	14.4014	3.4486	0.044	1.1193	0.0963	1.0692	0.3653	0.1880	0.1897	-0.2724
19	8.35	6.5568	1.7932	0.012	0.5698	0.0964	1.2153	0.1862	0.1537	0.0236	-0.0990
20	33.95	39.7380	-5.7880	0.132	-1.9967	0.1017	0.7598	-0.6718	0.1854	-0.2150	-0.0929
21	16.75	19.3642	-2.6142	0.051	-0.8731	0.1653	1.2377	-0.3885	-0.0841	-0.2972	0.3364
22	51.17	54.8565	-3.6865	0.451	-1.4896	0.3916	1.3981	-1.1950	0.5789	-1.0254	0.5731
23	17.6	22.2076	-4.6076	0.030	-1.4825	0.0413	0.8897	-0.3075	-0.1625	0.0373	-0.0527
24	18.68	23.2529	-4.5729	0.102	-1.5422	0.1206	0.9476	-0.5711	-0.2114	0.4046	-0.4654
25	9.6	9.8126	-0.2126	0.000	-0.0660	0.0666	1.2311	-0.0176	-0.0158	0.0008	0.0056

ตารางที่ 4.19: แสดงการตรวจสอบค่าสังเกตที่ผิดปกติและค่าสังเกตที่มีอิทธิพล

$$D_i > 1$$

$$|t_i| > 2$$

$$h_{ii} > \frac{2p}{n} = \frac{2(3)}{25} = 0.24 \quad \text{หรือ} \quad h_{ii} > \frac{3(3)}{25} = 0.36$$

$$DFFITS > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{3}{25}} = 0.69$$

$$DFBETAS > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{25}} = 0.40$$

$$COVRATIO > 1 - \frac{3p}{n} = 1 - \frac{3(3)}{25} = 0.64 \quad \text{และ} \quad COVRATIO < 1 + \frac{3p}{n} = 1 + \frac{3(3)}{25} = 1.36$$

จากตารางที่ 4.19 จะเห็นได้ว่าส่วนใหญ่แล้วค่าสังเกตที่ 9 และ 22 ให้ค่าที่ไม่สอดคล้องกับเกณฑ์ที่พิจารณา เมื่อพิจารณาค่า Leverage และ $DFFITS$ พบว่าค่าสังเกตที่ 9 และ 22 เป็นค่าสังเกตที่มีค่า Leverage สูงผิดปกติและมีอิทธิพลต่อค่าพยากรณ์ค่อนข้างมาก เมื่อพิจารณาค่า $DFBETAS$ พบว่าค่าสังเกตที่ 9 มีอิทธิพลอย่างมากต่อการประมาณระยะตัดแกน Y แต่มีอิทธิพลค่อนข้างน้อยต่อการประมาณ b_1 และ b_2 ในขณะที่ค่าสังเกตที่ 22 มีอิทธิพลอย่างมากต่อการประมาณ b_1 และเมื่อพิจารณาค่า $COVRATIO$ พบว่าการรวมเอาค่าสังเกตที่ 9 ไว้ จะทำให้ความแม่นยำของการประมาณลดลง ในขณะที่การรวมเอาค่าสังเกตที่ 22 จะช่วยเพิ่มความแม่นยำของการประมาณ กล่าวโดยสรุปจะได้ว่าค่าสังเกตที่ 9 และ 22 เป็นค่าสังเกตที่มีอิทธิพล

อย่างมากต่อสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด ดังนั้นจึงควรที่จะมีการตรวจสอบค่าสังเกตดังกล่าว หากไม่สามารถแก้ไขได้ อาจต้องใช้วิธีประมาณพารามิเตอร์ที่ถูกกระทบด้วยค่าสังเกตที่มีอิทธิพลหรือค่าที่ผิดปกติได้ค่อนข้างยาก

4.8 การตรวจสอบ Multicollinearity

ข้อกำหนดเบื้องต้นของการวิเคราะห์การถดถอยเชิงเส้นตรงแบบพหุกำหนดว่า ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระมีลักษณะเชิงเส้นตรง ความคลาดเคลื่อนเป็นอิสระกัน มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ σ^2 เท่ากัน นอกจากนี้ตัวแปรอิสระต้องไม่มีความสัมพันธ์ต่อกัน หากตัวแปรอิสระมีความสัมพันธ์ต่อกัน จะเรียกปัญหาที่เกิดขึ้นนี้ว่า *Intercorrelation* หรือ *Multicollinearity* โดยความสัมพันธ์ที่ใกล้เคียงเส้นตรง (Near-linear dependence) ระหว่างตัวแปรอิสระ จะทำให้เมตริกซ์ $\mathbf{X'X}$ เป็น Singular matrix และส่งผลกระทบต่อการประมาณค่าสัมประสิทธิ์ถดถอย

ปัญหา Multicollinearity จะส่งผลกระทบต่อการประมาณพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุดในโมเดลถดถอยเชิงเส้นตรงแบบพหุดังนี้

- เมื่อตัวแปรอิสระมีความสัมพันธ์ต่อกันสูง การประมาณค่าสัมประสิทธิ์ถดถอยมีแนวโน้มที่จะผันแปรสูง นั่นคือ S_{b_j} มีค่ามาก ทำให้การทดสอบอิทธิพลของตัวแปรอิสระแต่ละตัวต่อตัวแปรตาม Y พบว่าไม่มีนัยสำคัญทางสถิติ ในขณะที่การทดสอบนัยสำคัญของสมการระหว่างตัวแปรตาม Y และกลุ่มของตัวแปรอิสระ พบว่ามีนัยสำคัญทางสถิติ

- เมื่อเกิดปัญหา Multicollinearity การแปลผลค่าสัมประสิทธิ์ถดถอยว่าเป็นการวัดอัตราการเปลี่ยนแปลงของตัวแปรตามเฉลี่ยเมื่อตัวแปรอิสระที่พิจารณาเพิ่มขึ้น 1 หน่วย โดยที่ตัวแปรอิสระอื่นคงที่ ไม่สามารถใช้ได้อีกต่อไป เนื่องจากค่าสัมประสิทธิ์ถดถอยของตัวแปรอิสระตัวใดตัวหนึ่งจะขึ้นอยู่กับตัวแปรอิสระอื่นทั้งที่รวมอยู่ในโมเดลหรืออาจไม่ได้รวมอยู่ในโมเดล

- เมื่อตัวแปรอิสระมีความสัมพันธ์ต่อกันสูง การอธิบายความผันแปรทั้งหมดที่ลดลงเนื่องจากตัวแปรอิสระตัวใดตัวหนึ่ง ไม่สามารถทำได้อีกต่อไป เพราะอิทธิพลของตัวแปรดังกล่าวจะขึ้นอยู่กับตัวแปรอิสระอื่นที่สัมพันธ์กันที่รวมอยู่ในโมเดล ซึ่งจะส่งผลกระทบต่อค่าสัมประสิทธิ์ตัวกำหนดบางส่วนด้วยเช่นกัน นอกจากนี้ค่า Extra-sum-of-squares ของตัวแปรอิสระดังกล่าวเมื่อมีตัวแปรอิสระอื่นที่สัมพันธ์กันรวมอยู่ในโมเดล อาจมีค่าไม่น้อยกว่า Sum-of-squares ของตัวแปรอิสระนั้นเดี่ยว ๆ เช่น ถ้า X_1 และ X_2 มีความสัมพันธ์กันสูง อาจได้ว่า $SSR(X_2|X_1) > SSR(X_2)$ โดยจะเรียกตัวแปร X_1 ว่า *Suppressor variable*

- ปัญหา Multicollinearity ไม่มีผลกระทบต่อการพยากรณ์ รวมทั้งการสร้างช่วงแห่งความเชื่อมั่นของค่าเฉลี่ยของตัวแปรตามและช่วงแห่งการพยากรณ์ของค่าสังเกตตัวใหม่ ถ้าการอนุมานดังกล่าวยังคงอยู่ในขอบเขตของค่าสังเกตที่ใช้ในการสร้างสมการ

จะเห็นได้ว่าปัญหา Multicollinearity มีผลกระทบต่อการประมาณค่าสัมประสิทธิ์ถดถอยอย่างมาก ซึ่งอาจทำให้ผลการวิเคราะห์เบี่ยงเบนไปจากที่ควรจะเป็นได้ ดังนั้นการตรวจสอบว่าเกิดปัญหา Multicollinearity หรือไม่ จึงเป็นสิ่งสำคัญที่ควรคำนึงถึง โดยสัญญาณที่อาจชี้ว่าเกิดปัญหา Multicollinearity มีดังนี้

- พิจารณาค่าสัมประสิทธิ์ถดถอยว่ามีการเปลี่ยนแปลงอย่างมากหรือไม่ เมื่อมีการเพิ่มตัวแปรอิสระตัวใดตัวหนึ่งเข้าไปในโมเดลหรือตัดตัวแปรอิสระบางตัวออกจากโมเดล รวมทั้งเมื่อมีการเปลี่ยนแปลงค่าสังเกตหรือตัดค่าสังเกตบางตัวออกจากการวิเคราะห์
- การทดสอบอิทธิพลของตัวแปรอิสระที่สำคัญ พบว่าตัวแปรดังกล่าวมีอิทธิพลต่อตัวแปรตามอย่างไม่มีนัยสำคัญทางสถิติ
- ค่าสัมประสิทธิ์ถดถอยมีเครื่องหมายตรงข้ามกับที่คาดไว้ตามทฤษฎีหรือประสบการณ์เดิม
- ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคู่ของตัวแปรอิสระมีค่าค่อนข้างสูง
- ช่วงความเชื่อมั่นของสัมประสิทธิ์ถดถอยของตัวแปรอิสระที่สำคัญบางตัวมีช่วงที่ค่อนข้างกว้าง

นอกจากนี้การตรวจสอบ Multicollinearity ยังสามารถทำได้โดยคำนวณค่า *Variance Inflation Factor* หรือนิยมเรียกย่อ ๆ ว่า *VIF* ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$VIF_j = \frac{1}{1 - R_j^2} \tag{4.68}$$

เมื่อ

VIF_j แทน ค่า Variance Inflation Factor ของตัวแปรอิสระตัวที่ j

R_j^2 แทน ค่าสัมประสิทธิ์ตัวกำหนดระหว่างตัวแปรอิสระ X_j กับตัวแปรอิสระอื่น

จะเห็นได้ว่าถ้า X_j ไม่มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระอื่น R_j^2 มีค่าเข้าใกล้ 0 และทำให้ VIF_j มีค่าเข้าใกล้ 1 แต่ถ้า X_j มีความสัมพันธ์เกือบจะเป็นเส้นตรงกับตัวแปรอิสระอื่น R_j^2 มีค่าเข้าใกล้ 1 ส่งผลให้ VIF_j มีค่าค่อนข้างมาก ดังนั้นค่า VIF_j ที่มากกว่า 1 มาก ๆ อาจชี้ว่าเกิดปัญหา Multicollinearity ได้ โดยทั่วไปจะถือว่าค่า $VIF_j > 5$ หรือ 10 แสดงถึงการเกิดปัญหา Multicollinearity และถ้า $VIF_j > 10$ แสดงถึงการเกิดปัญหา Multicollinearity อย่างรุนแรง

วิธีแก้ไขเมื่อเกิดปัญหา Multicollinearity

- เก็บรวบรวมข้อมูลเพิ่มเติม เพื่อลดความรุนแรงของ Multicollinearity แต่อย่างไรก็ตามการเพิ่มขนาดตัวอย่าง อาจไม่สามารถทำได้เสมอไป เนื่องจากข้อจำกัดทางด้านงบประมาณ หรือกระบวนการที่ศึกษาเสร็จสิ้นไปแล้ว ทำให้ไม่สามารถเก็บตัวอย่างเพิ่มได้ ถึงแม้ว่าบางครั้งจะเก็บข้อมูลเพิ่มเติมได้ แต่ข้อมูลที่ได้อาจไม่เหมาะสม เนื่องจากไม่อยู่ในขอบเขตที่พิจารณา หรือถ้าข้อมูลใหม่มีลักษณะที่ผิดปกติ ก็อาจมีอิทธิพลอย่างมากต่อโมเดล ดังนั้นจะเห็นได้ว่าการเพิ่มขนาดตัวอย่างจึงไม่ใช่วิธีการแก้ปัญหา Multicollinearity ที่ชัดเจน

- ระบุรูปแบบของโมเดลใหม่ เพื่อลดอิทธิพลของ Multicollinearity ซึ่งถ้า X_1 และ X_2 มีความสัมพันธ์เชิงเส้นตรงต่อกันสูง อาจใช้ฟังก์ชันอื่นของตัวแปรอิสระ เช่น $X = X_1X_2$ เป็นต้น ที่ยังคงให้ข้อมูลเช่นเดิม แต่ช่วยลดปัญหา Multicollinearity หรืออีกวิธีหนึ่งที่นิยมใช้กันอย่างแพร่หลาย คือ การตัดตัวแปรอิสระบางตัว เช่น อาจตัดตัวแปรอิสระ X_2 ออกจากโมเดลที่มีตัวแปร X_1 และ X_2 ที่มีความสัมพันธ์เชิงเส้นตรงต่อกันสูง ซึ่งการตัดตัวแปรอิสระมักเป็นวิธีที่มีประสิทธิภาพในการลดปัญหา Multicollinearity แต่อาจทำให้ความสามารถในการพยากรณ์ของโมเดลลดลง ดังนั้นการเลือกตัวแปรอิสระจะต้องทำด้วยความระมัดระวัง

- ประมาณค่าพารามิเตอร์ของสมการถดถอยด้วยวิธี Ridge regression เนื่องจากปัญหา Multicollinearity ทำให้ตัวประมาณค่าสัมประสิทธิ์ถดถอยที่ได้ด้วยวิธีกำลังสองน้อยที่สุดยังคงเป็นตัวประมาณที่ไม่เอนเอียง แต่ไม่มีความแปรปรวนต่ำสุด ดังนั้นวิธี Ridge regression จะบรรเทาปัญหาดังกล่าวด้วยการค้นหาตัวประมาณที่เอนเอียง แต่มีความแปรปรวนต่ำกว่าตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุด

- ประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธี Principal components regression ซึ่งได้จากการสร้างตัวแปรอิสระตัวใหม่ที่เป็นอิสระกัน เรียกว่า *Principal components* แล้วจึงประมาณพารามิเตอร์โดยใช้เพียงบางเซ็ตของ Principal components ที่สร้างขึ้น

- ประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธี Latent root regression ซึ่งมีหลักการคล้ายคลึงกับวิธี Principal components โดยหาตัวประมาณจากค่า Eigenvalues และ Eigenvectors ของเมตริกซ์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม

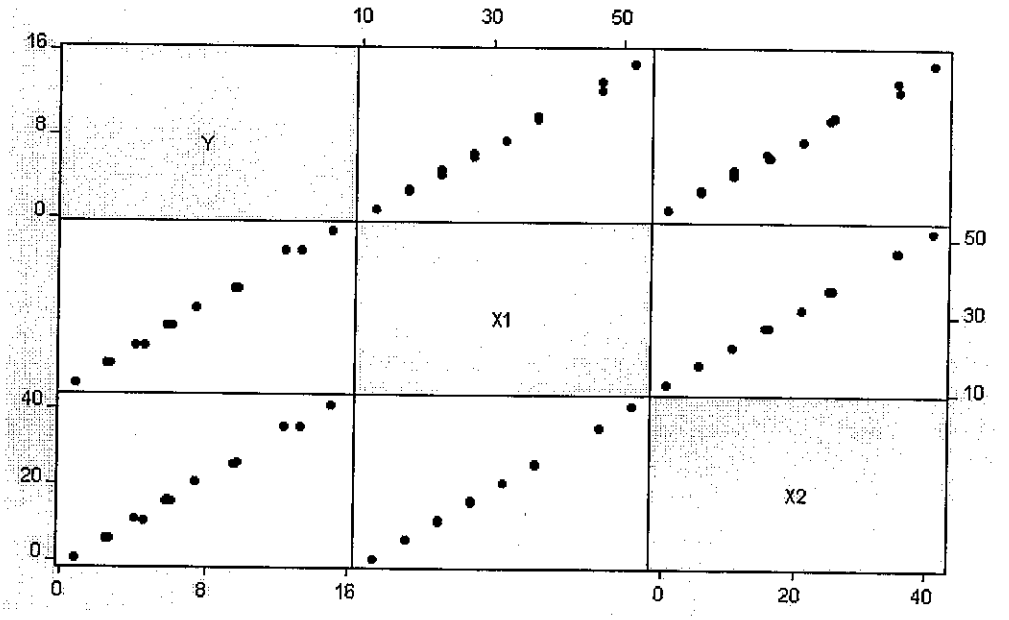
$$\begin{bmatrix} 1 & \mathbf{Y'X} \\ \mathbf{X'Y} & \mathbf{X'X} \end{bmatrix}$$

วิธี Latent root regression เป็นวิธีหนึ่งที่ช่วยขจัดปัญหา Multicollinearity แต่มีประโยชน์ต่อการพยากรณ์ค่อนข้างน้อย

ตัวอย่างที่ 4.12 จากข้อมูลต่อไปนี้ จงตรวจสอบว่าเกิดปัญหา Multicollinearity หรือไม่

Y	X_1	X_2
13.23	47	35.38
0.76	12	0.43
5.84	27	15.96
9.55	37	25.25
7.35	32	20.84
2.69	17	5.36
12.35	47	35.58
4.61	22	10.46
14.95	52	40.97
9.65	37	25.63
5.80	27	15.64
2.45	17	5.48
9.75	37	25.78
6.10	27	15.31
4.06	22	10.58

ตารางที่ 4.20: ข้อมูลสำหรับตัวอย่างที่ 4.12



รูปที่ 4.18: Matrix plot ระหว่างตัวแปรทุกคู่

วิธีทำ จากข้อมูลในตาราง 4.21 สร้างแผนภาพการกระจายระหว่างตัวแปรทุกคู่ดังแสดงในรูปที่ 4.18 ซึ่งจะเห็นได้ว่าตัวแปรอิสระ X_1 และ X_2 มีความสัมพันธ์เชิงเส้นตรงกับ Y ค่อนข้างสูง และในขณะเดียวกันมีความสัมพันธ์เชิงเส้นตรงต่อกันค่อนข้างสูงด้วย ซึ่งสอดคล้องกับค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_1 และ X_2 โดยที่ $r_{12} = 0.999$ ดังแสดงในตาราง 4.21 และชี้ว่าเกิดปัญหา Multicollinearity ค่อนข้างรุนแรง นอกจากนี้ผลที่ได้จากโปรแกรม MINITAB ให้ค่า $VIF_1 = VIF_2 = 3151.8$ แสดงว่าเกิดปัญหา Multicollinearity อย่างรุนแรงเช่นเดียวกัน

	Y	X_1	X_2
Y	1.000	0.998	0.997
X_1		1.000	0.999
X_2			1.000

ตารางที่ 4.21: เมตริกซ์ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลในตัวอย่างที่ 4.12

สร้างสมการถดถอยของข้อมูลข้างต้นได้เป็น

$$\hat{Y} = \underset{(4.120)}{-6.8930} + \underset{(0.3566)}{0.6500}X_1 - \underset{(0.3544)}{-0.2996}X_2$$

โดยที่ $R^2 = 0.996$ และ $MSE = 0.0816$

เนื่องจาก X_1 และ X_2 มีความสัมพันธ์ต่อกันค่อนข้างสูง จึงอาจเลือกตัดตัวแปรตัวใดตัวหนึ่งออกจากสมการ หากตัด X_2 ออกจากสมการข้างต้น จะได้ว่า

$$\hat{Y} = \underset{(0.2060)}{-3.4134} + \underset{(0.0063)}{0.3486}X_1$$

โดยที่ $R^2 = 0.996$ และ $MSE = 0.0798$

หากตัด X_1 ออกจากสมการข้างต้น จะได้ว่า

$$\hat{Y} = \underset{(0.1543)}{0.6135} + \underset{(0.0069)}{0.3462}X_2$$

โดยที่ $R^2 = 0.995$ และ $MSE = 0.0962$

จะเห็นได้ว่าสมการที่ตัด X_2 ออกไป มีค่า R^2 ที่สูงกว่า รวมทั้งมีค่า MSE ที่ต่ำกว่าสมการที่ตัด X_1 ออกไป ดังนั้นการจะรับเอาสมการใดไปใช้ ควรที่จะมีการวิเคราะห์ความคลาดเคลื่อนและตรวจสอบความเหมาะสมของรูปแบบโมเดลก่อน

ข้อสังเกต เมื่อเทียบกับสมการที่มีตัวแปรอิสระเพียงตัวเดียวกับสมการที่ตัวแปรอิสระ 2 ตัว จะเห็นได้ว่าค่า S_{b_1} และ S_{b_2} ของสมการที่มีตัวแปรอิสระ 2 ตัว มีค่าเพิ่มขึ้นอย่างมาก ซึ่งเป็นสัญญาณหนึ่งที่ชี้ว่าเกิดปัญหา Multicollinearity

แบบฝึกหัดบทที่ 4

1. นักเคมีคนหนึ่งต้องการศึกษาความสัมพันธ์ระหว่างการรวมตัวของสารละลาย (Y) และระยะเวลา (X) โดยรวบรวมข้อมูลจากการทดลอง 12 ครั้ง ได้ผลดังนี้

ค่าสังเกตที่	1	2	3	4	5	6	7	8	9	10	11	12
X	10	6	8	4	6	4	4	2	4	2	1	2
Y	0.7	1.7	1.5	4.7	5.9	5.5	12.0	11.4	10.8	28.2	25.8	30.8

1.1 จงสร้างแผนภาพการกระจาย และอธิบายรูปแบบความสัมพันธ์ของตัวแปรทั้งสองจากกราฟ

1.2 จงสร้างสมการถดถอยเชิงเส้นตรงอย่างง่าย พร้อมทั้งคำนวณค่าความคลาดเคลื่อน (Residuals)

1.3 จงตรวจสอบว่ารูปแบบสมการถดถอยเชิงเส้นตรงที่ใช้เหมาะสมกับข้อมูลหรือไม่

1.4 จงสร้างกราฟของความคลาดเคลื่อน (Residual plots) เพื่อตรวจสอบว่าความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่หรือไม่

1.5 จงตรวจสอบว่าข้อมูลที่ใช้มีการแจกแจงแบบปกติหรือไม่

1.6 จงตรวจสอบว่าข้อมูลชุดนี้มีค่าสังเกตที่ผิดปกติ (Outliers) หรือค่าสังเกตที่มีอิทธิพล (Influential points) เกิดขึ้นหรือไม่

1.7 ที่ระดับนัยสำคัญ 0.01 จงแสดงการทดสอบ Lack of fit และสรุปผลที่ได้

2. บริษัทผลิตไมโครคอมพิวเตอร์แห่งหนึ่งต้องการศึกษาความสัมพันธ์ระหว่างจำนวนสินค้าที่ผลิต (Y_t) หน่วยเป็นพันล้านบาท และมูลค่าที่ใช้ในการผลิต (X_t) หน่วยเป็นล้านบาท เป็นเวลา 16 เดือน ได้ข้อมูลดังนี้

เดือนที่	1	2	3	4	5	6	7	8	9	10
X_t	2.05	2.03	2.00	1.95	1.94	1.89	1.99	2.05	2.10	2.11
Y_t	102.9	101.5	100.8	98.0	97.3	93.5	97.5	102.2	105.0	107.2
เดือนที่	11	12	13	14	15	16				
X_t	2.06	2.06	2.04	2.08	2.10	2.15				
Y_t	105.1	105.9	103.0	104.8	105.0	107.2				

2.1 จงสร้างสมการถดถอยเชิงเส้นตรงอย่างง่าย พร้อมทั้งคำนวณค่าความคลาดเคลื่อน (Residuals)

2.2 จงสร้างกราฟของความคลาดเคลื่อนกับลำดับเวลา ท่านคิดว่าเกิดปัญหา Autocorrelation กับข้อมูลชุดนี้หรือไม่ ถ้าเกิดปัญหาดังกล่าว ท่านคิดว่าเป็น Positive หรือ Negative autocorrelation

2.3 ที่ระดับนัยสำคัญ 0.05 จงแสดงการทดสอบสมมติฐานเพื่อตรวจสอบว่าความคลาดเคลื่อนเกิดปัญหาของ Positive autocorrelation หรือไม่ โดยใช้สถิติทดสอบของ Durbin-Watson

2.4 ถ้าผลสรุปจากข้อ 2.3 ระบุว่าความคลาดเคลื่อนมี Autocorrelation เกิดขึ้น จงแก้ปัญหาดังกล่าวโดยการแปลงข้อมูล ใช้วิธีของ Cochrane Orcutt ในการประมาณค่า Population autocorrelation (ρ) และตรวจสอบว่าการแปลงข้อมูลที่ใช้สามารถขจัดปัญหาของการเกิด Autocorrelation ได้หรือไม่ พร้อมทั้งแปลงสมการถดถอยที่ได้ให้อยู่ในรูปของตัวแปรเดิม

3. ในการศึกษาความสัมพันธ์ระหว่างระดับความอ้วน (Y) กับตัวแปรอิสระ 3 ตัว คือ ความหนาของชั้นผิวหนัง (X_1) เส้นรอบวงของต้นขา (X_2) และเส้นรอบวงของกึ่งกลางแขน (X_3) ได้ข้อมูลดังนี้

Y	X_1	X_2	X_3
12.4	20.7	42.4	29.7
0.5	25.9	49.1	28.8
19.2	31.9	51.2	37.6
20.6	31.0	53.6	31.7
13.4	20.3	41.5	31.5
22.2	26.8	53.2	24.3
27.6	32.6	57.8	28.2
25.9	29.1	51.4	31.2
21.8	23.3	49.2	23.8
19.8	26.7	52.8	25.4
25.9	32.3	55.9	30.6
27.7	31.6	56.0	28.9
12.2	19.9	45.8	23.6
18.3	20.9	43.5	29.2
13.3	15.8	42.0	21.9
24.4	30.7	53.7	30.7
23.1	28.9	54.6	26.3
25.9	31.4	57.9	25.2
15.3	23.9	47.5	27.7
21.6	26.4	50.3	28.1

3.1 จงสร้างแผนภาพการกระจายของตัวแปรแต่ละคู่ และคำนวณเมตริกซ์ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation matrix) ของตัวแปรแต่ละคู่ ท่านคิดว่าข้อมูลชุดนี้เกิดปัญหา Multicollinearity หรือไม่ ทราบได้อย่างไร

3.2 จงสร้างสมการถดถอยเชิงเส้นตรงแบบพหุแสดงความสัมพันธ์ระหว่างตัวแปรตาม Y กับตัวแปรอิสระทั้ง 3 ตัว พร้อมทั้งคำนวณค่า VIF (Variance Inflation Factor) และท่านคิดว่าเกิดปัญหา Multicollinearity หรือไม่ ทราบได้อย่างไร

3.3 หากเกิดปัญหา Multicollinearity กับข้อมูลชุดนี้ ท่านจะแก้ปัญหาดังกล่าวอย่างไร แสดงวิธีทำด้วย

4. ในการศึกษาความสัมพันธ์ระหว่างระดับของเม็ดเลือดแดง (Y) และอายุ (X) ของเด็กกลุ่มหนึ่ง ได้ข้อมูลดังนี้

ค่าสังเกตที่	1	2	3	4	5	6	7	8	9	10	11	12	13
X	0	0	0	0	0	1	1	1	1	1	2	2	2
Y	13.44	12.84	11.91	15.60	20.09	10.11	11.38	10.28	8.96	8.59	9.83	9.00	8.65
ค่าสังเกตที่	14	15	16	17	18	19	20	21	22	23	24	25	
X	2	2	3	3	3	3	3	4	4	4	4	4	
Y	7.85	8.88	7.94	6.01	5.14	6.90	6.77	4.86	5.10	5.67	5.75	6.23	

4.1 จงสร้างแผนภาพการกระจายระหว่างตัวแปรตาม Y และตัวแปรอิสระ X จากกราฟท่านคิดว่ารูปแบบความสัมพันธ์เป็นเส้นตรงหรือไม่

4.2 ที่ระดับนัยสำคัญ 0.01 จงทดสอบ Lack of fit ของสมการถดถอยเชิงเส้นตรง และสรุปผลที่ได้

4.3 จงแปลงข้อมูลของ Y ให้อยู่ในรูปของ $\ln Y$ และสร้างสมการถดถอยเชิงเส้นตรง โดยให้ตัวแปรตามอยู่ในรูปของ $\ln Y$ พร้อมทั้งทดสอบ Lack of fit ของสมการถดถอยเชิงเส้นตรงอีกครั้งที่ระดับนัยสำคัญ 0.01 และสรุปผลที่ได้

4.4 จงแปลงข้อมูลของตัวแปรตาม Y โดยใช้ Box-Cox transformation เพื่อหากำลัง (Power) ของการแปลงข้อมูลที่เหมาะสม กำหนดให้ λ มีค่าตั้งแต่ -2 ถึง 1 เพิ่มครั้งละ 0.01 ท่านจะเลือกกำลังของการแปลงข้อมูลเป็นเท่าไร