บทที่ 3
การวิเคราะห์ความเหมาะสมของตัวแบบ

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายมีประโยชน์และมีการใช้กันอย่างแพร่หลายในงานวิจัยแต่การที่จะใช้ตัวแบบการถดถอยให้มีความถูกต้องและมีอำนาจการทดสอบที่สูงนั้น ตัวแบบจำเป็นต้องเป็นไปตามข้อตกลง (assumption) ข้อตกลงที่กำหนดไว้มี 5 ข้อ (Montgomery & Peck, 1992, p. 7) ดังนี้คือ

1. ความสัมพันธ์ระหว่างตัวแปร X และ Y เป็นเส้นตรงหรือมีแนวโน้มเป็นเส้นตรง
2. ความคลาดเคลื่อน (ε) มีค่าเฉลี่ยเท่ากับ 0
3. ความคลาดเคลื่อน (ε) มีความแปรปรวนคงที่เท่ากับ σ^2
4. ความคลาดเคลื่อนแต่ละค่าเป็นอิสระต่อกัน
5. ความคลาดเคลื่อนมีการแจกแจงแบบปกติ

ข้อตกลงดังกล่าวมีความจำเป็นอย่างยิ่งโดยเฉพาะในการทดสอบสมมติฐานและการประมาณค่าพารามิเตอร์ต่างๆ หากตัวแบบไม่เป็นไปตามข้อตกลงแล้วอาจทำให้การตัดสินใจผิดพลาดได้ เนื่องจากค่าประมาณของความคลาดเคลื่อนคือส่วนเหลือ (residual หรือ e) ซึ่งส่วนเหลือสามารถคำนวณได้ดังนี้

$$e_i = Y_i - \hat{Y}_i \tag{3.1}$$

หากตัวแบบการถดถอยที่ได้ไม่เป็นไปตามข้อตกลงแล้วจะพบว่าส่วนเหลือจะเบี่ยงเบนไปจากข้อตกลง ดังนั้นการวิเคราะห์ส่วนเหลือจึงเป็นการวิเคราะห์ตัวแบบที่ง่ายและมีประโยชน์มาก

3.1 คุณสมบัติของส่วนเหลือ

เนื่องจากข้อตกลงของการใช้สมการถดถอยกล่าวว่าความคลาดเคลื่อนมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่และความคลาดเคลื่อนแต่ละค่าเป็นอิสระต่อกัน ดังนั้นส่วนเหลือมีค่าเฉลี่ยเท่ากับ 0 หรือ

$$\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0 \tag{3.2}$$

และความแปรปรวนของส่วนเหลือคือ

$$\frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n-2}=\frac{\sum_{i=1}^{n}e_i^2}{n-2}=\frac{SSE}{n-2}=MSE \tag{3.3}$$

เนื่องจากองศาเสรีเท่ากับ $n-2$ แสดงว่าส่วนเหลือไม่เป็นอิสระต่อกันทั้งนี้เนื่องจากส่วนเหลือมาจากค่าพยากรณ์ ($\hat{Y}$) และค่าพยากรณ์มีการประมาณค่า 2 ค่าคือ b_0 และ b_1 ดังนั้นส่วนเหลือจึงสูญเสียความเป็นอิสระไป 2 ค่าเช่นเดียวกับค่าพยากรณ์ Montgomery และ Peck (1992) หน้า 68 กล่าวว่าหากข้อมูลมีขนาดใหญ่พอแล้วตัวแบบจะได้รับผลกระทบเพียงเล็กน้อยจากการที่ส่วนเหลือไม่เป็นอิสระต่อกัน และหากจำนวนข้อมูลมีขนาดใหญ่มากเมื่อเทียบกับจำนวนตัวแปรอิสระแล้วตัวแบบจะได้รับผลกระทบเพียงเล็กน้อยเช่นกัน เมื่อค่าความแปรปรวนของส่วนเหลือเพิ่มมากขึ้นความถูกต้องในการพยากรณ์จะลดลง (Freund, Wilson, & Sa, 2006, p. 43)

ในบางกรณีส่วนเหลือมาตรฐาน (standardized residual) จะมีประโยชน์อย่างมากในการวิเคราะห์และการตีความหมายโดยเฉพาะอย่างยิ่งในการวิเคราะห์การถดถอยพหุ ส่วนเหลือมาตรฐานได้จากการหารส่วนเหลือด้วยค่าส่วนเบี่ยงเบนมาตรฐานทำให้ส่วนเหลือเป็นค่าที่ไม่มีหน่วยโดยมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1

$$e_i^*=\frac{e_i}{\sqrt{MSE}} \tag{3.4}$$

ส่วนเหลือปรับแล้ว (studentized residual) เป็นอีกรูปแบบของส่วนเหลือที่มีความสำคัญและใช้กันมากในการวิเคราะห์ตัวแบบ ส่วนเหลือปรับแล้วจะหารด้วยความแปรปรวนของส่วนเหลือแต่ละค่าในขณะที่ส่วนเหลือมาตรฐานจะหารด้วยความแปรปรวนรวม

$$e_i'=\frac{e_i}{\sqrt{MSE\left[1-\left(\frac{1}{n}+\frac{(X_i-\bar{x})^2}{S_{xx}}\right)\right]}} \tag{3.5}$$

การใช้ข้อมูลจำนวนน้อยนั้นจะทำให้ส่วนเหลือแต่ละค่ามีความแปรปรวนที่ต่างกันมากการหารด้วยความแปรปรวนรวมจึงอาจไม่เหมาะสมดังนั้นการใช้ส่วนเหลือปรับแล้วจะเหมาสมกว่าการใช้ส่วนเหลือมาตรฐาน แต่เมื่อขนาดตัวอย่างใหญ่ขึ้นความแปรปรวนของส่วนเหลือแต่ละค่าจะ

ไม่แตกต่างกันมากดังนั้นการใช้ส่วนเหลือมาตรฐานหรือส่วนเหลือปรับแล้วจะให้ผลที่ไม่แตกต่างกันมาก

3.2 การวิเคราะห์ส่วนเหลือ

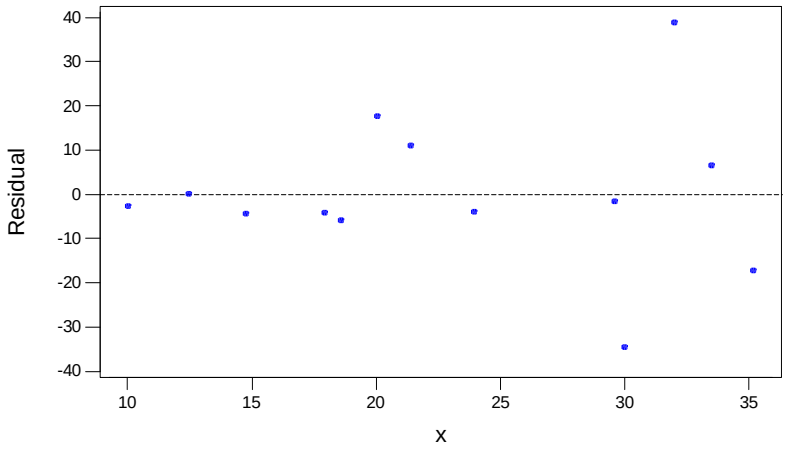
การวิเคราะห์ส่วนเหลือสามารถทำได้ 6 แบบหลักคือ

1. ฟังก์ชันการถดถอยไม่เป็นเส้นตรง
2. ส่วนเหลือมีความแปรปรวนที่ไม่คงที่
3. ส่วนเหลือไม่เป็นอิสระต่อกัน
4. การมีค่าผิดปกติ (outlier)
5. ส่วนเหลือมีการแจกแจงที่ไม่เป็นปกติ
6. ตัวแปรอิสระที่สำคัญบางตัวไม่ถูกรวมเข้าในการสร้างตัวแบบถดถอย

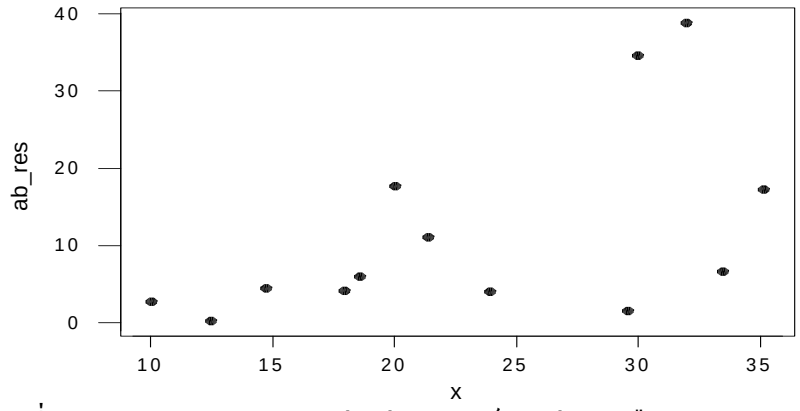
การตรวจสอบข้อตกลงของความคลาดเคลื่อนจึงทำได้โดยการตรวจสอบจากส่วนเหลือ การตรวจสอบข้อตกลงอย่างง่ายคือการวาดแผนภาพกระจายของส่วนเหลือและพิจารณาว่าเป็นไปตามข้อตกลงหรือไม่ แต่บางครั้งแผนภาพกระจายที่ได้อาจไม่ชัดเจนเพียงพอที่จะใช้ในการพิจารณาได้ การวาดแผนภาพกระจายเพื่อตรวจสอบ 6 รูปแบบดังกล่าวสามารถวาดไว้ 7 ลักษณะดังนี้

1. แผนภาพกระจายระหว่างส่วนเหลือกับตัวแปรอิสระ (ภาพที่ 3.1)
2. แผนภาพกระจายระหว่างค่าสัมบูรณ์ของส่วนเหลือหรือค่าส่วนเหลือกำลังสองกับตัวแปรอิสระ (ภาพที่ 3.2)
3. แผนภาพกระจายระหว่างส่วนเหลือกับค่าพยากรณ์ (ภาพที่ 3.3)
4. แผนภาพกระจายระหว่างส่วนเหลือกับเวลา (ภาพที่ 3.4)
5. แผนภาพกระจายระหว่างส่วนเหลือกับตัวแปรอิสระที่ไม่ได้รวมอยู่ในตัวแบบ
6. แผนภาพกล่อง (box plot) ของส่วนเหลือ (ภาพที่ 3.5)
7. แผนภาพกระจายแบบปกติ (normality probability plot) ของส่วนเหลือ (ภาพที่ 3.6)

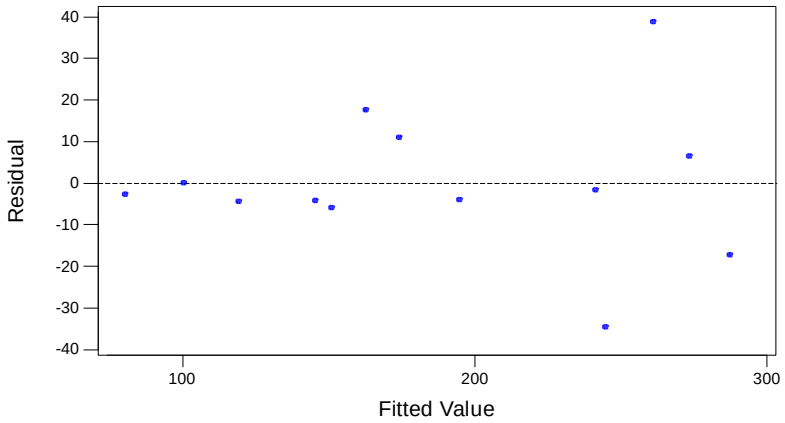
ภาพที่ 3.1 – 3.6 ได้จากข้อมูลในตัวอย่างที่ 1.1 ที่วาดโดยใช้โปรแกรม MINITAB



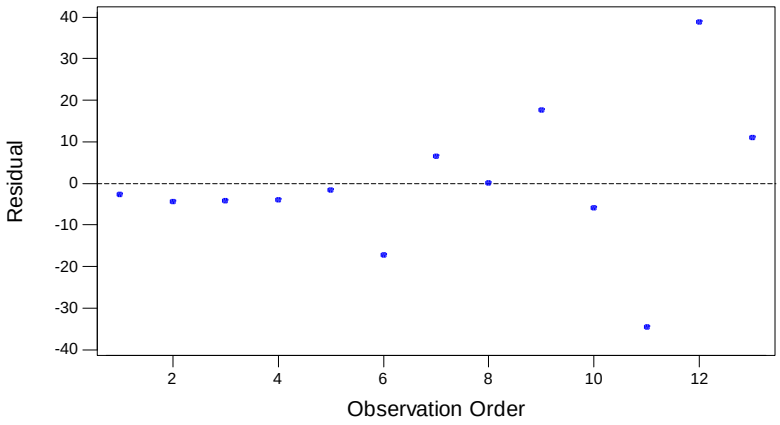
ภาพที่ 3.1 แผนภาพกระจายระหว่างส่วนเหลือกับตัวแปรอิสระ



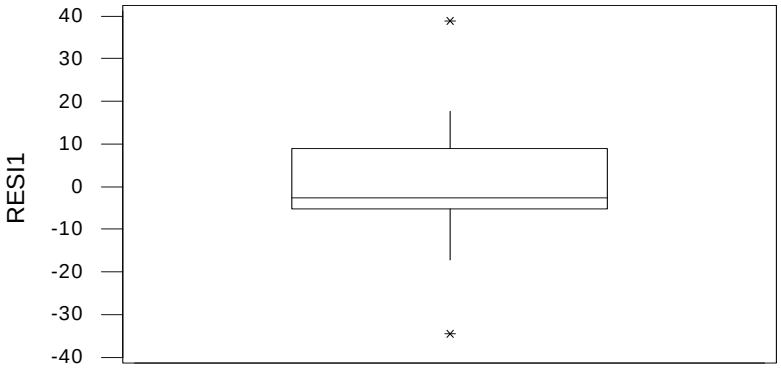
ภาพที่ 3.2 แผนภาพกระจายระหว่างค่าสัมบูรณ์ของส่วนเหลือกับตัวแปรอิสระ



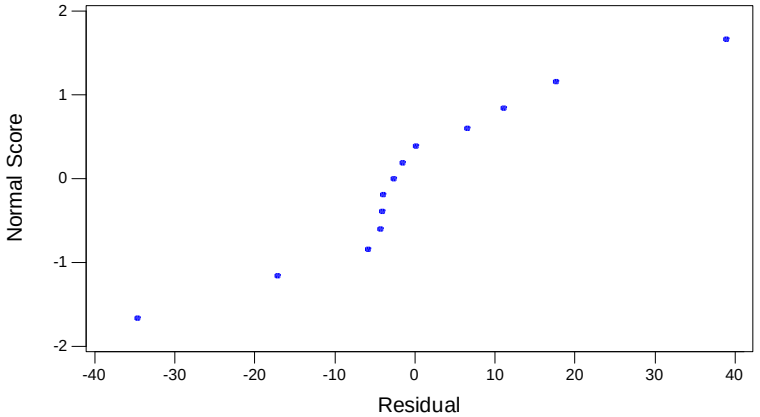
ภาพที่ 3.3 แผนภาพกระจายระหว่างส่วนเหลือกับค่าพยากรณ์



ภาพที่ 3.4 แผนภาพกระจายระหว่างส่วนเหลือกับเวลา



ภาพที่ 3.5 แผนภาพกล่องของส่วนเหลือ

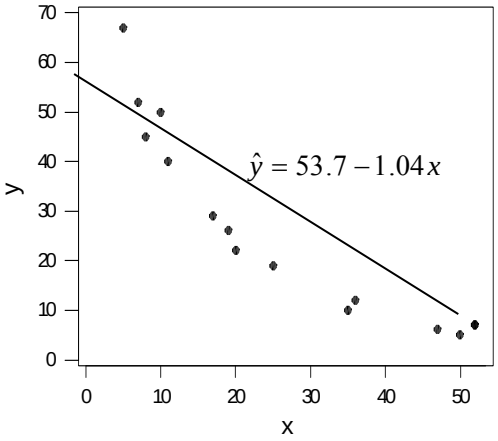


ภาพที่ 3.6 แผนภาพกระจายแบบปกติของส่วนเหลือ

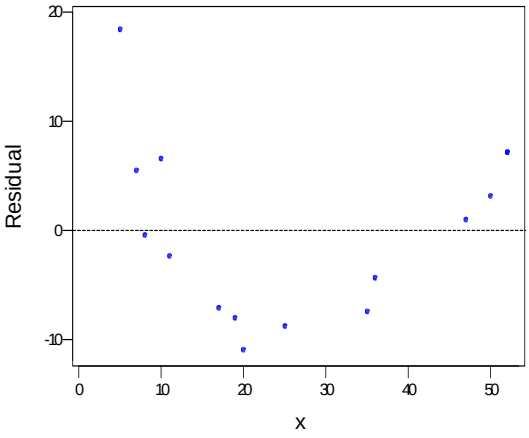
3.2.1 ฟังก์ชันการถดถอยไม่เป็นเส้นตรง

ในการทดสอบความเป็นฟังก์ชันเชิงเส้นตรงสามารถทำได้โดยการวาดแผนภาพกระจายระหว่างตัวแปรอิสระและตัวแปรตามแต่วิธีนี้อาจยากต่อการสังเกต แผนภาพกระจายระหว่างส่วนเหลือกับตัวแปรอิสระหรือระหว่างส่วนเหลือกับค่าพยากรณ์จะง่ายแก่การวิเคราะห์มากกว่า หากฟังก์ชันการถดถอยเป็นเส้นตรงแล้วส่วนเหลือจะกระจายรอบค่าศูนย์ในแนวนอนหรือกระจายอย่างไม่เป็นระบบหรือไม่สามารถหารูปแบบได้ ในกรณีของสมการถดถอยเชิงเส้นอย่างง่ายแล้วแผนภาพกระจายของตัวแปรอิสระกับส่วนเหลือจะให้ผลเหมือนกับแผนภาพกระจายของค่าพยากรณ์กับส่วนเหลือ

เมื่อพิจารณาแผนภาพกระจายระหว่างตัวแปร X และ Y ที่มีความสัมพันธ์เชิงเส้นโค้งกันดังภาพ 3.7 ก พบว่าจุดต่างๆ มีแนวโน้มเป็นเส้นโค้งและการกระจายระหว่างตัวแปร X กับส่วนเหลือในภาพ 3.7 ข จะมีลักษณะโค้งโดยส่วนเหลือจะมีค่าบวกเมื่อ X มีค่าน้อยและเมื่อ X มีค่ามากแต่จะมีค่าลบเมื่อ X มีค่ากลาง ดังนั้นการใช้สมการถดถอยเชิงเส้นอย่างง่ายจึงไม่เหมาะกับข้อมูลชนิดนี้



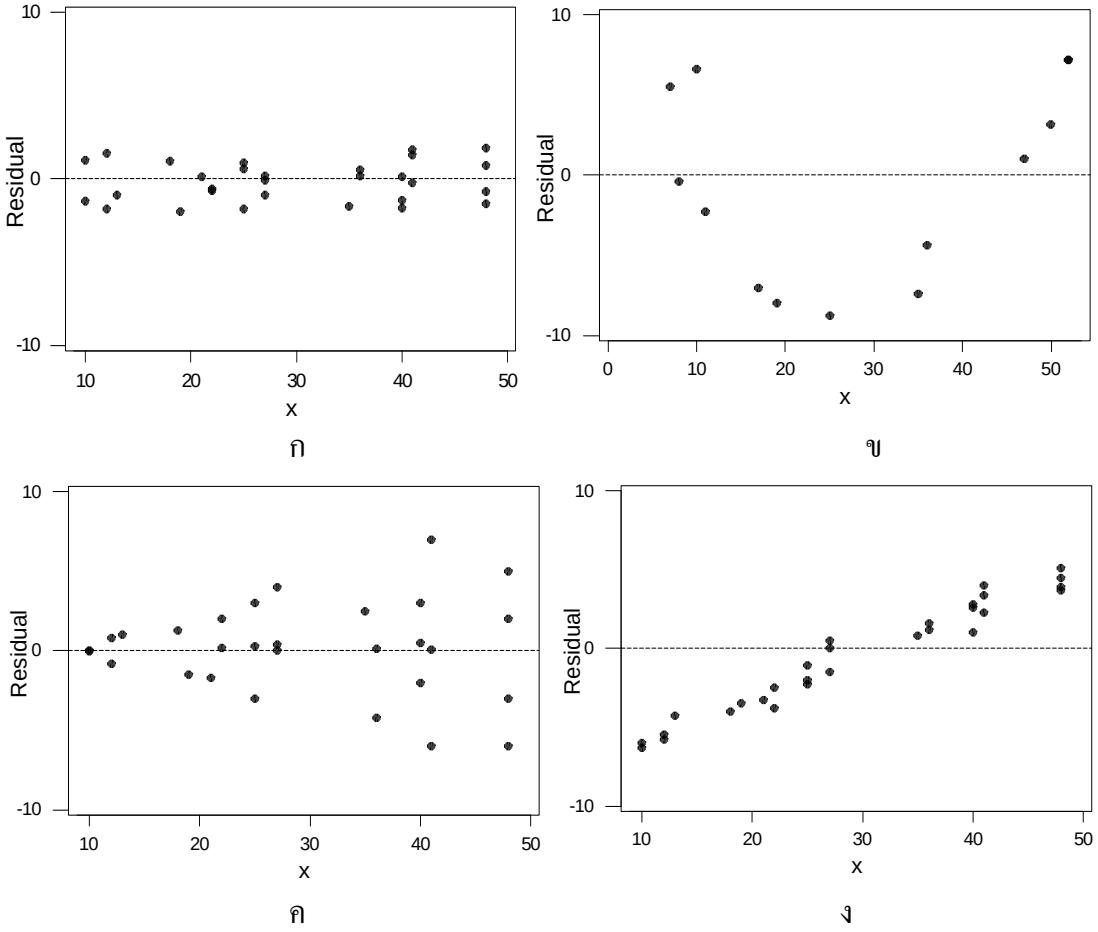
ภาพที่ 3.7 ก การกระจายของ X กับ Y



ภาพที่ 3.7 ข การกระจายของ X กับส่วนเหลือ

ภาพที่ 3.8 ก – ง แสดงให้เห็นถึงความสัมพันธ์ของตัวแปรอิสระกับส่วนเหลือในรูปแบบต่างๆ กัน โดยภาพที่ 3.8 ก เป็นรูปของความสัมพันธ์เชิงเส้นตรงของตัวแปรอิสระและตัวแปรตามโดยส่วนเหลือมีการกระจายตัวที่สม่ำเสมอรอบค่า 0 สำหรับภาพที่ 3.8 ข ค และ ง แสดงให้เห็นถึงความสัมพันธ์ระหว่างตัวแปรทั้งสองที่ไม่เป็นเส้นตรง โดยภาพที่ 3.8 ข แผนภาพกระจายมีลักษณะเป็นเส้นโค้งพาราโบลา ภาพที่ 3.8 ค แผนภาพกระจายมีลักษณะเป็นรูปปากแตรกล่าวคือ เมื่อ X มีค่าต่ำแล้วความแปรปรวนของส่วนเหลือจะน้อยและความแปรปรวนของส่วนเหลือเพิ่มขึ้นเมื่อ X มีค่าเพิ่มขึ้น

ภาพที่ 3.8 ง แผนภาพกระจายมีแนวโน้มเพิ่มมากขึ้น เมื่อ X มีค่าต่ำแล้วส่วนเหลือจะมีค่าเป็นลบและส่วนเหลือจะมีค่าเป็นบวกเพิ่มมากขึ้นเมื่อ X มีค่าเพิ่มขึ้น



ภาพที่ 3.8 การกระจายของตัวแปรอิสระกับส่วนเหลือในรูปแบบต่างๆ

3.2.2 ส่วนเหลือมีความแปรปรวนที่ไม่คงที่

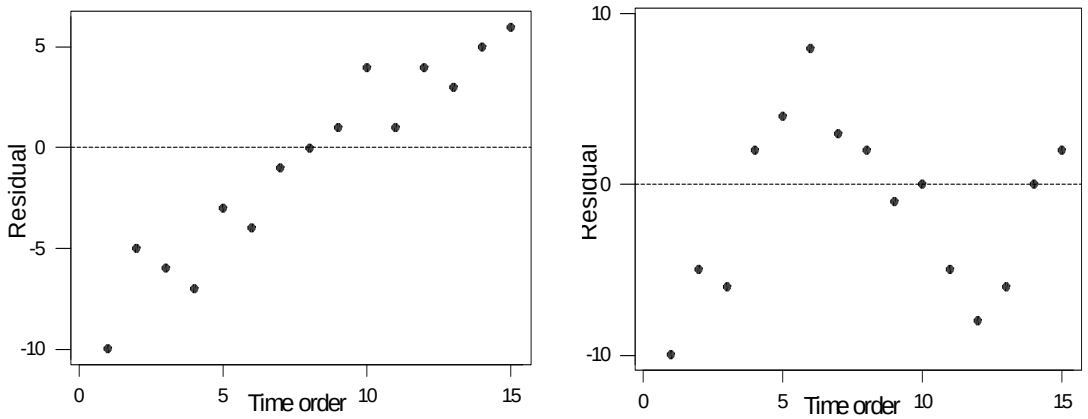
การวิเคราะห์ว่าส่วนเหลือมีความแปรปรวนคงที่หรือไม่สามารถพิจารณาจากแผนภาพต่างๆ ดังนี้ แผนภาพกระจายระหว่างตัวแปรอิสระกับส่วนเหลือ แผนภาพกระจายระหว่างค่าพยากรณ์กับส่วนเหลือ แผนภาพกระจายระหว่างตัวแปรอิสระกับค่าสัมบูรณ์ของส่วนเหลือหรือกับส่วนเหลือกำลังสอง แผนภาพกระจายระหว่างค่าพยากรณ์กับค่าสัมบูรณ์ของส่วนเหลือกับส่วนเหลือกำลังสอง ในภาพที่ 3.8 ก ส่วนเหลือมีการกระจายตัวที่สม่ำเสมอแสดงว่าส่วนเหลือมีความแปรปรวนที่คงที่ เมื่อส่วนเหลือมีความแปรปรวนไม่คงที่สามารถตรวจสอบได้จากแผนภาพกระจายระหว่างตัวแปรอิสระกับส่วนเหลือดังในภาพที่ 3.8 ค ลักษณะของความแปรปรวนของส่วนเหลือที่เพิ่มมากขึ้นเมื่อตัวแปรอิสระมีค่ามากขึ้นจะพบได้บ่อย ลักษณะเช่นนี้เรียก megaphone

ตัวอย่างเช่น เมื่อเด็กหญิงที่มีอายุเพิ่มมากขึ้น ความดันเลือดจะมีความแปรปรวนเพิ่มขึ้น เป็นต้น (Neter et al., 1996, p. 102)

แผนภาพกระจายระหว่างตัวแปรอิสระหรือค่าพยากรณ์กับค่าสัมบูรณ์ของส่วนเหลือหรือกับส่วนเหลือกำลังสองนั้นจะมีประโยชน์มากโดยเฉพาะอย่างยิ่งในกรณีที่ข้อมูลมีจำนวนไม่มากเนื่องจากค่าสัมบูรณ์หรือค่ากำลังสองนี้จะทำให้เห็นความไม่คงที่ของความแปรปรวนอย่างชัดเจน

3.2.3 ส่วนเหลือไม่เป็นอิสระต่อกัน

หากข้อมูลที่นำมาวิเคราะห์มีความเกี่ยวข้องกับเวลา เช่น การวัดอุณหภูมิของน้ำในทะเลสาบทุกชั่วโมง เป็นต้น ข้อมูลชนิดนี้จะไม่เป็นอิสระต่อกัน วิธีการตรวจสอบอย่างง่ายคือการวาดแผนภาพกระจายระหว่างส่วนเหลือกับเวลา หากพบว่าส่วนเหลือมีการกระจายที่มีแบบแผน เช่น มีแนวโน้มที่สูงขึ้นหรือลดลงดังภาพที่ 3.9 ก หรือเป็นวัฏจักรดังภาพที่ 3.9 ข เป็นต้น แสดงว่าข้อมูลชุดนี้ไม่เป็นอิสระต่อกัน หากข้อมูลเป็นอิสระต่อกันแล้วแผนภาพที่ได้นั้นส่วนเหลือควรมีการกระจายที่สม่ำเสมอรอบค่าศูนย์



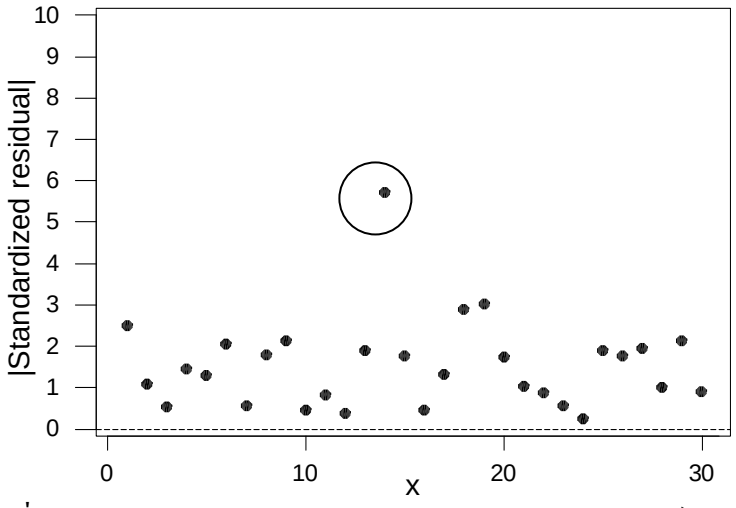
ก ข

ภาพที่ 3.9 การกระจายของเวลากับส่วนเหลือในรูปแบบต่างๆ

3.2.4 การมีค่าผิดปกติ

หากข้อมูลมีค่าผิดปกติ (outlier) แล้วจะทำให้ส่วนเหลือมีค่ามากกว่าหรือน้อยกว่าค่าส่วนเหลืออื่นๆ ตามไปด้วยแต่หากค่าผิดปกติมีมากกว่า 1 ค่าและมีค่าใกล้เคียงกันแล้วบางครั้งจะยากแก่การตรวจสอบ การตรวจสอบค่าผิดปกติอย่างง่ายคือ การวาดแผนภาพกระจายระหว่างตัวแปรอิสระกับส่วนเหลือหรือค่าพยากรณ์กับส่วนเหลือ นอกจากนี้ยังสามารถวาดแผนภาพกล่อง

(box plot) แผนภาพก้านและใบ (stem-and-leaf plot) และแผนภาพความน่าจะเป็นของการแจกแจงแบบปกติ (normal probability plot) ของส่วนเหลือเพื่อตรวจสอบค่าผิดปกติก็ได้ วิธีตรวจสอบที่มีประสิทธิภาพวิธีหนึ่งคือการวาดแผนภาพกระจายของตัวแปรอิสระกับส่วนเหลือมาตรฐาน หากข้อมูลใดมีค่าสัมบูรณ์ของส่วนเหลือมาตรฐานที่มากกว่าหรือเท่ากับ 4 แล้วข้อมูลนั้นเป็นค่าผิดปกติ เพราะโอกาสที่ข้อมูลจะมีค่าสัมบูรณ์มากกว่า 3 (หรือมากกว่า $\pm$ 3 ของส่วนเบี่ยงเบนมาตรฐาน) นั้นมีน้อยมาก (Neter et al., 1996, p. 103 และ Montgomery & Peck, 1992, p. 80) ค่าผิดปกติจะมีผลต่อความถูกต้องของสมการถดถอยและค่าพยากรณ์อย่างมากเนื่องจากสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุดนั้นจะไวต่อค่าผิดปกติดังนั้นจึงจำเป็นอย่างยิ่งที่จะตรวจสอบค่าผิดปกติโดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีจำนวนไม่มาก ภาพที่ 3.10 แสดงให้เห็นมีข้อมูลที่มีค่ามากกว่า 4 อยู่ 1 ค่าดังนั้นข้อมูลนี้อาจเป็นค่าผิดปกติ



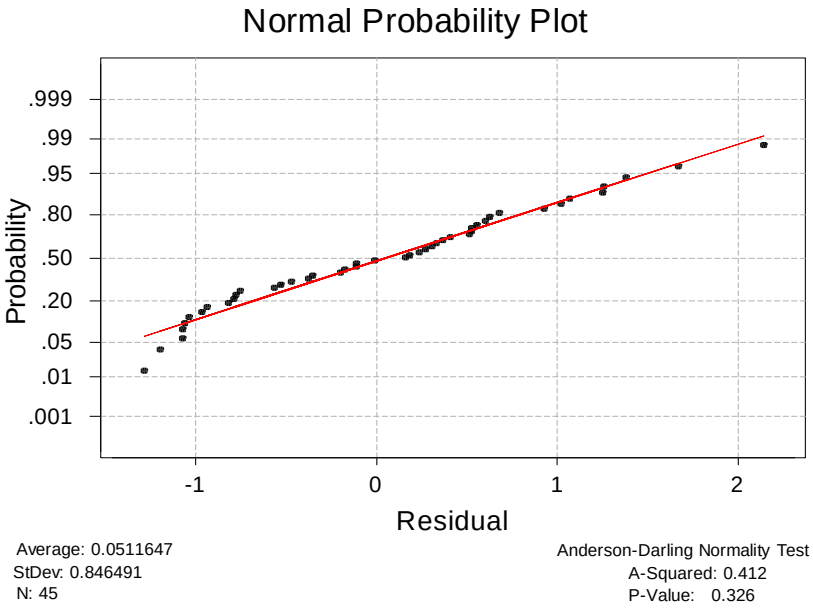
ภาพที่ 3.10 แผนภาพกระจายระหว่างตัวแปรอิสระกับค่าสัมบูรณ์ของส่วนเหลือมาตรฐาน

ค่าผิดปกติบางครั้งอาจทำให้ส่วนเหลือไม่เป็นไปตามข้อตกลงที่กำหนดไว้หากตัดค่าผิดปกตินี้ออกไปอาจทำให้ส่วนเหลือเป็นไปตามข้อตกลงก็ได้ นอกจากนี้ค่าผิดปกติยังทำให้ค่าสัมประสิทธิ์การตัดสินใจมีค่าสูงขึ้นและค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนมีค่าลดลงด้วย

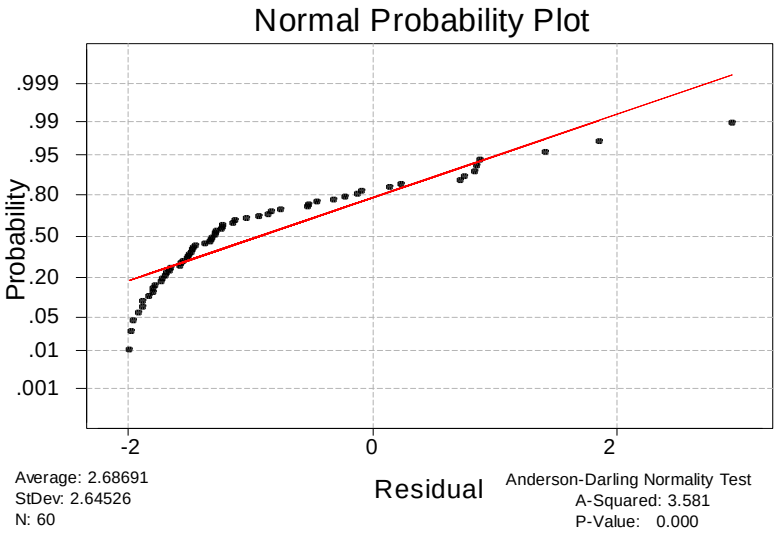
3.2.5 ส่วนเหลือมีการแจกแจงที่ไม่เป็นปกติ

หากส่วนเหลือมีการแจกแจงที่ไม่เป็นปกติเพียงเล็กน้อยจะไม่มีผลกระทบต่อการวิเคราะห์โดยใช้วิธีการกำลังสองน้อยที่สุด แต่หากส่วนเหลือมีการแจกแจงที่เบี่ยงเบนจากการแจกแจงปกติมากแล้วอาจทำให้ผลการวิเคราะห์ผิดพลาดได้ การพิจารณาว่าส่วนเหลือมีการแจกแจง

ปกติหรือไม่อย่างง่ายคือ การใช้แผนภูมิและแผนภาพที่มีด้วยกันหลายประเภท เช่น แผนภูมิฮีสโตแกรม แผนภาพกล่อง แผนภาพแบบจุด (dot plot) และแผนภาพความน่าจะเป็นของการแจกแจงแบบปกติ เป็นต้น ในการใช้แผนภาพความน่าจะเป็นของการแจกแจงแบบปกติเพื่อพิจารณานั้นหากจุดต่างๆ เรียงตัวกันในเชิงเส้นตรง 45 องศาแล้วส่วนเหลือนั้นมีการแจกแจงแบบปกติ ภาพที่ 3.11 แสดงการกระจายตัวในแบบต่างๆ โดยการใช้แผนภาพความน่าจะเป็นของการแจกแจงแบบปกติที่ใช้โปรแกรม MINITAB วาด โดยรูป ก แสดงให้เห็นถึงการแจกแจงของส่วนเหลือที่เป็นปกติโดยค่าส่วนใหญ่อยู่ในแนวเส้นตรง 45 องศานอกจากนี้ค่า *p*-value ที่ได้ยังมีค่ามากกว่า 0.05 โดยใช้วิธีทดสอบการแจกแจงแบบปกติของ Anderson – Darling แสดงว่าส่วนเหลือชุดนี้มีการแจกแจงแบบปกติ สำหรับภาพที่ 3.11 ข นั้นส่วนเหลือมีการแจกแจงแบบเบ้ซ้ายโดยจะเห็นว่าจุดต่างๆ มีลักษณะโค้งที่ส่วนเหลือที่มีค่าน้อยมีความน่าจะเป็นต่ำกว่าที่ควรจะเป็นเช่นเดียวกับส่วนเหลือที่มีค่ามาก แต่ในขณะที่ส่วนเหลือที่มีค่ากลางจะมีความน่าจะเป็นที่สูงเกินกว่าที่ควรจะเป็นซึ่งสอดคล้องกับค่า *p*-value ที่ต่ำ



ก



ข

ภาพที่ 3.11 แผนภาพการทดสอบความน่าจะเป็นของการแจกแจงแบบปกติ

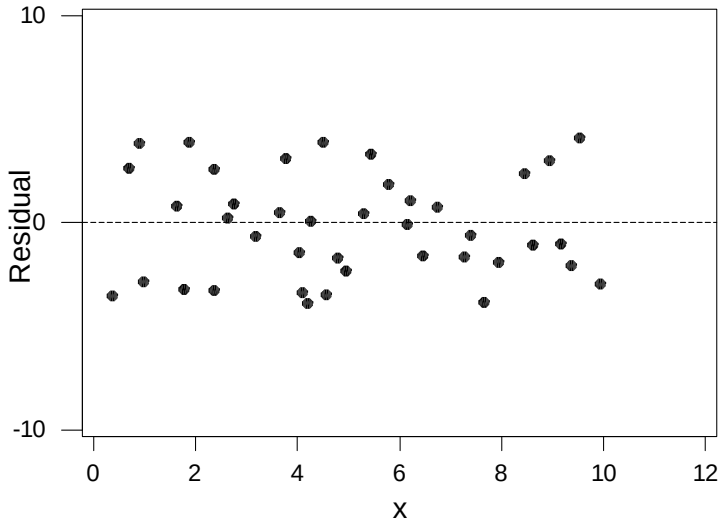
การตรวจสอบการแจกแจงแบบปกติควรทำการตรวจสอบหลังจากการตรวจสอบและแก้ไขตามข้อ 1 - 4 เสียก่อนเนื่องจากว่าบางครั้งการที่แผนภาพแสดงให้เห็นว่าส่วนเหลือไม่มีการแจกแจงแบบปกตินั้นอาจมาจากการที่ส่วนเหลือมีความแปรปรวนที่ไม่คงที่แต่เมื่อมีการแก้ไขแล้วส่วนเหลืออาจมีการแจกแจงที่ปกติก็ได้

หากตัวอย่างมีขนาดใหญ่แล้วการที่ส่วนเหลือไม่มีการแจกแจงแบบปกติจะมีผลกระทบต่อค่าพยากรณ์ของสัมประสิทธิ์เพียงเล็กน้อย (Mooi & Sarstedt, 2011, p. 174-175)

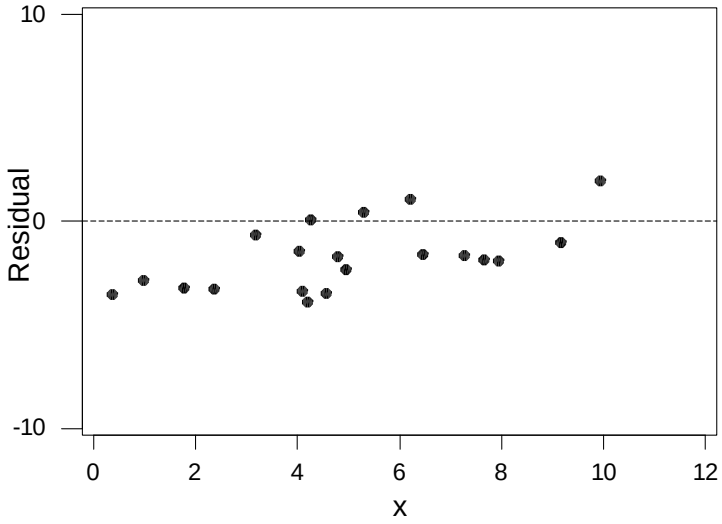
3.2.6 ตัวแปรอิสระที่สำคัญบางตัวไม่ถูกรวมเข้าในการสร้างตัวแบบถดถอย

การที่ตัวแปรอิสระที่สำคัญบางตัวถูกละเลยหรือไม่ได้นำมารวมในการสร้างตัวแบบจะทำให้ผลการวิเคราะห์ที่ได้มีความผิดพลาดและอาจทำให้ผลรวมกำลังสองความคลาดเคลื่อน (*SSE*) มีค่าสูงเกินกว่าที่ควรจะเป็นและส่งผลให้ไม่ปฏิเสธสมมติฐานในการวิเคราะห์ความแปรปรวนได้ การที่จะทราบว่าตัวแปรอิสระตัวใดถูกละเลยนั้นต้องมาจากการศึกษาจากทฤษฎีหรือจากงานวิจัยที่เกี่ยวข้องมาช่วยในการวินิจฉัย การวิเคราะห์ว่าตัวแปรอิสระตัวใดถูกละเลยจากตัวแบบโดยวิธีใช้แผนภาพนั้นทำได้โดยการวาดแผนภาพกระจายระหว่างตัวแปรอิสระนั้นกับส่วนเหลือว่ามีรูปแบบที่เปลี่ยนไปหรือไม่ เช่น ในการศึกษาความเร็วในการประกอบชิ้นส่วนกับอายุงานของพนักงานแห่งหนึ่งมีแผนภาพกระจายของอายุงานกับส่วนเหลือดังภาพที่ 3.12 ก จากนั้นหากพิจารณาแผนภาพแยกตามเพศจะพบว่าเพศชายส่วนใหญ่จะมีค่าส่วนเหลือที่ต่ำกว่าค่าศูนย์หรือเพศชายใช้เวลาในการประกอบจริงต่ำกว่าที่ตัวแบบพยากรณ์ได้ดังภาพที่ 3.12 ข ในขณะที่เพศหญิง

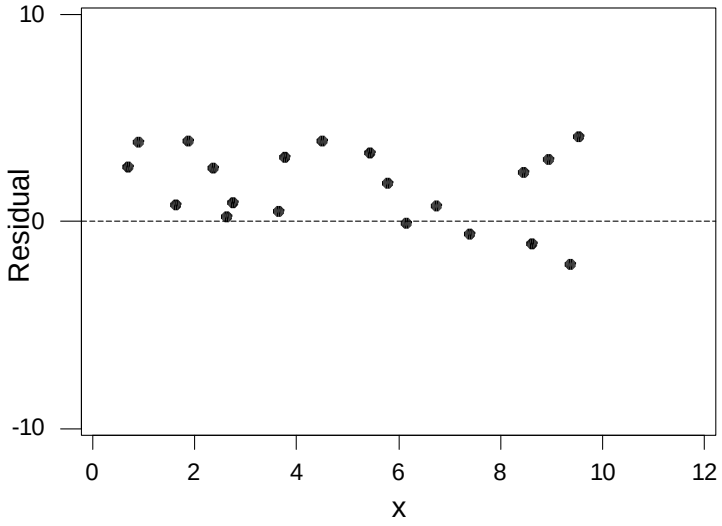
ส่วนใหญ่จะมีค่าส่วนเหลือที่สูงกว่าค่าศูนย์หรือเพศหญิงใช้เวลาในการประกอบจริงสูงกว่าที่ตัวแบบพยากรณ์ได้ดังภาพที่ 3.12 ค แสดงว่าเพศทั้งสองมีความแตกต่างในการทำงานดังนั้นการสร้างตัวแบบโดยไม่คำนึงถึงเพศจะทำให้การพยากรณ์เวลาที่ใช้ผิดไปจากความเป็นจริงจึงควรเพิ่มตัวแปรเพศเข้าในการสร้างตัวแบบเพิ่มความถูกต้องของตัวแบบ



ก แผนภาพที่ใช้ข้อมูลทั้งสองเพศ



ข แผนภาพที่ใช้ข้อมูลเฉพาะเพศชาย



ค แผนภาพที่ใช้ข้อมูลเฉพาะเพศหญิง

ภาพที่ 3.12 แผนภาพกระจายระหว่างอายุงานกับเวลาที่ใช้

3.2.7 การวิเคราะห์ส่วนเหลือโดยใช้การทดสอบทางสถิติ

ถึงแม้ว่าการใช้กราฟหรือแผนภาพในการวิเคราะห์ส่วนเหลือนั้นง่ายแก่การพิจารณา แต่บางครั้งอาจยากแก่การตัดสินใจโดยเฉพาะอย่างยิ่งเมื่อกราฟหรือแผนภาพนั้นแสดงภาพที่ไม่ชัดเจน ดังนั้นการใช้วิธีทดสอบทางสถิติจึงจำเป็นในกรณีที่ ไม่สามารถตัดสินใจได้ การทดสอบทางสถิติให้ค่าที่เชื่อถือได้มากกว่า ข้อเสียของการทดสอบทางสถิติคือ การทดสอบส่วนมากมีข้อตกลงว่าข้อมูลหรือส่วนเหลือต้องไม่มีความสัมพันธ์กันหรือเป็นอิสระต่อกันแต่หากข้อมูลมีขนาดใหญ่เพียงพอแล้วสามารถยอมให้ข้อมูลมีความสัมพันธ์กันได้ (Neter et al., 1996, p. 110) การทดสอบทางสถิติของส่วนเหลือที่จะกล่าวถึงคือ การทดสอบความเป็นอิสระ การทดสอบความคงที่ของความแปรปรวน การทดสอบค่าผิดปกติและการทดสอบการกระจายตัวแบบปกติ

3.2.7.1 การทดสอบความเป็นอิสระของส่วนเหลือ การทดสอบความเป็นอิสระของส่วนเหลือที่นิยมใช้คือวิธี Durbin-Watson test การทดสอบนี้จะทดสอบว่าส่วนเหลือในช่วงเวลาปัจจุบันมีความสัมพันธ์กับส่วนเหลือในช่วงเวลา 1 ช่วงก่อนหน้านี้หรือไม่ ส่วนใหญ่แล้วหากส่วนเหลือมีความสัมพันธ์กันมักจะเป็นทางบวกขั้นตอนการทดสอบมีดังนี้

(1) การตั้งสมมติฐาน

$H_0 : \rho \leq 0$ หรือ ส่วนเหลือไม่มีความสัมพันธ์กันทางบวก

$H_1 : \rho > 0$ หรือ ส่วนเหลือมีความสัมพันธ์กันทางบวก

(2) กำหนดระดับนัยสำคัญ

(3) กำหนดเขตวิกฤต

สถิติที่ใช้ในการทดสอบนี้ คือ D และเนื่องจากค่าวิกฤตของการทดสอบนี้ยากแก่การคำนวณดังนั้น Durbin และ Watson จึงใช้เขตวิกฤตที่มีการใช้ค่าต่ำสุด (d_L) และค่าสูงสุด (d_U) เขตวิกฤตนี้ขึ้นกับจำนวนตัวแปรอิสระที่ใช้

(4) คำนวณค่าสถิติ

$$D = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (3.6)$$

ค่าสถิตินี้ได้จากการเปรียบเทียบค่าส่วนเหลือในปัจจุบันกับค่าในอดีต 1 ช่วงเวลาก่อนหน้า

(5) สรุปผล

การสรุปผลว่าส่วนเหลือมีความสัมพันธ์กันหรือไม่นั้นพิจารณาดังนี้

หาก $D > d_U$ แล้วสรุปว่าส่วนเหลือไม่มีความสัมพันธ์กันทางบวก

หาก $D < d_L$ แล้วสรุปว่าส่วนเหลือมีความสัมพันธ์กันทางบวก

หาก $d_L \leq D \leq d_U$ แล้วไม่สามารถสรุปได้ว่ามีความสัมพันธ์กันทางบวกหรือไม่

หมายเหตุ

1. หากต้องการทดสอบความสัมพันธ์เชิงลบหรือ $H_1 : \rho < 0$ แล้วสถิติ D จะเป็น $4 - D$ หาก $4 - D$ มีค่าน้อยกว่า d_L แล้วสรุปว่าส่วนเหลือมีความสัมพันธ์กันทางลบ
2. หากต้องการทดสอบทั้งสองด้านในเวลาเดียวกันหรือ $H_1 : \rho \neq 0$ แล้วให้ใช้ระดับนัยสำคัญเป็น 2α เมื่อ α คือระดับนัยสำคัญสำหรับการทดสอบด้านเดียว
3. หากค่าสถิติ D ตกอยู่ในเขตที่ไม่สามารถตัดสินใจได้แล้วควรทำการเก็บข้อมูลเพิ่ม
4. การทดสอบนี้ใช้ในการทดสอบส่วนเหลือที่มีความสัมพันธ์กับช่วงเวลาก่อนหน้าเท่านั้นไม่สามารถใช้กับความสัมพันธ์แบบอื่นได้

ตัวอย่างที่ 3.1 ข้อมูลของปริมาณปุ๋ยที่ใช้กับความสูงของต้นไม้จงทดสอบความสัมพันธ์ทางบวกของส่วนเหลือโดยใช้วิธี Durbin-Watson test ที่ระดับนัยสำคัญ 0.05

ปริมาณ (X_i)	ความสูง (Y_i)	e_i	$e_i - e_{i-1}$	$(e_i - e_{i-1})^2$	$(e_i)^2$
5.6	66	-5.235	-	-	27.40523
3.0	71	2.322	7.557	57.1082	5.391684
4.5	78	7.687	5.365	28.7832	59.08997
2.5	67	-1.705	-9.392	88.2097	2.907025
5.2	61	-10.123	-8.418	70.8627	102.4751
5.8	75	4.292	14.415	207.7922	18.42126
4.0	65	-4.747	-9.039	81.7035	22.53401
3.8	79	9.680	14.427	208.1383	93.7024
6.1	76	4.326	-5.354	28.6653	18.71428
4.7	67	-3.357	-7.683	59.0285	11.26945
5.4	69	-2.341	1.016	1.0323	5.480281
4.1	66	-3.404	-1.063	1.1300	11.58722
6.3	72	0.317	3.721	13.8458	0.100489
4.4	77	7.146	6.829	46.6352	51.06532
3.2	64	-4.859	-12.005	144.1200	23.60988
รวม			**0.376**	**1037.055**	**453.7536**

จากการสร้างสมการถดถอยโดยใช้วิธีกำลังสองน้อยที่สุดได้ค่าส่วนเหลือดังตารางข้างบน เนื่องจากระดับนัยสำคัญเท่ากับ 0.05 และจำนวนตัวแปรอิสระเท่ากับ 1 ดังนั้นเขตวิกฤตที่ได้จากตารางที่ 3 ในภาคผนวก คือ d_L เท่ากับ1.08 และ d_U เท่ากับ 1.36 ค่าสถิติสามารถคำนวณได้ดังนี้

$$D = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

$$D = \frac{1037.055}{453.7536} = 2.285$$

พบว่าค่าสถิติ D (2.285) > d_U (1.36) ดังนั้นส่วนเหลือไม่มีความสัมพันธ์ทางบวกกัน

3.2.7.2 การทดสอบความคงที่ของความแปรปรวน การทดสอบความคงที่ของความแปรปรวนที่จะกล่าวถึงคือวิธี Modified Levene test การทดสอบนี้เป็นการทดสอบว่าความแปรปรวนของส่วนเหลือเพิ่มขึ้นตามค่าของตัวแปรอิสระที่เพิ่มขึ้นหรือลดลงตามค่าของตัวแปรอิสระที่เพิ่มขึ้นหรือไม่ดังภาพที่ 3.8 ค ส่วนเหลือไม่จำเป็นต้องมีการแจกแจงแบบปกติแต่วิธีนี้ขนาดตัวอย่างต้องมีขนาดใหญ่พอเพื่อที่จะตัดผลกระทบจากการที่ส่วนเหลือไม่เป็นอิสระต่อกัน เมื่อส่วนเหลือมีความแปรปรวนสูงขึ้นจะพบว่าส่วนเหลือจะมีความแปรผันมากขึ้นด้วย ดังนั้นการตรวจสอบของวิธีนี้จะแบ่งส่วนเหลือออกเป็นสองส่วนตามค่าของตัวแปรอิสระจากนั้นเปรียบเทียบค่าเฉลี่ยของความแตกต่างสัมบูรณ์ของส่วนเหลือกับค่ามัธยฐานของแต่ละกลุ่มทั้งสองกลุ่มว่าต่างกันหรือไม่โดยใช้สถิติ t ขั้นตอนการทดสอบมีดังนี้

(1) การตั้งสมมติฐาน

H_0 : ความแปรปรวนของส่วนเหลือคงที่

H_1 : ความแปรปรวนของส่วนเหลือไม่คงที่

(2) กำหนดระดับนัยสำคัญ

(3) กำหนดเขตวิกฤต

เขตวิกฤตสามารถหาได้จากการอ่านค่า t ที่องศาเสรี $n-2$ และ $\alpha/2$

(4) คำนวณค่าสถิติ

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{3.7}$$

โดย $n = n_1 + n_2$

$d_{i1} = \left|e_{i1} - \tilde{e}_1\right|$

$d_{i2} = \left|e_{i2} - \tilde{e}_2\right|$

$\tilde{e}_1$ = มัธยฐานของส่วนเหลือในกลุ่มที่ 1

$\tilde{e}_2$ = มัธยฐานของส่วนเหลือในกลุ่มที่ 2

$$s = \frac{\sum_{i=1}^{n_1}\left(d_{i1} - \tilde{d}_1\right)^2 + \sum_{i=1}^{n_2}\left(d_{i2} - \tilde{d}_2\right)^2}{n-2}$$

(5) สรุปผล

หากค่า $\left|t_L^*\right| \le t_{n-2,\alpha/2}$ แล้วสรุปว่าความแปรปรวนของส่วนเหลือคงที่

หากค่า $\left|t_L^*\right| > t_{n-2,\alpha/2}$ แล้วสรุปว่าความแปรปรวนของส่วนเหลือไม่คงที่

หมายเหตุ

หากระดับค่าของตัวแปรอิสระมีหลายกลุ่มแล้วให้ทดสอบโดยใช้การทดสอบ *t* ของสองกลุ่มประชากรทำการทดสอบของ 2 กลุ่มใหญ่โดยรวมกลุ่มต่างๆ เข้ากับกลุ่มที่มีค่าของตัวแปรอิสระสูงสุด

ตัวอย่างที่ 3.2 จากข้อมูลในตัวอย่าง 3.1 จงใช้วิธี Modified Levene test ในการทดสอบความคงที่ของความแปรปรวนของส่วนเหลือที่ระดับนัยสำคัญ 0.05

วิธีทำ

เขตวิกฤตสามารถหาได้จากตารางที่ 1 ในภาคผนวกโดยมีค่า *t* ที่องศาเสรีเท่ากับ $15-2=13$ และ $\alpha=0.05$ มีค่าเท่ากับ 2.160 เนื่องจากค่าของตัวแปรอิสระมีการกระจายที่สม่ำเสมอตั้งแต่ 2.5 จนถึง 6.3 ดังนั้นจะแบ่งส่วนเหลือเป็น 2 กลุ่มดังนี้ กลุ่มแรกมีจำนวน 8 ตัวตั้งแต่ค่า 2.5 จนถึง 4.5 และกลุ่มที่สองมีจำนวน 7 ตัวตั้งแต่ค่า 4.7 จนถึง 6.3 จะได้

$\tilde{e}_1 = 0.3085$ และ $\tilde{e}_2 = -2.34$

จากนั้นคำนวณค่า d_{i1} และ d_{i2} ดังนี้

$$d_{11} = |e_{11} - \tilde{e}_1| = |-1.705 - 0.3085| = 2.0135$$

$$d_{12} = |e_{12} - \tilde{e}_2| = |-3.357 - (-0.2341)| = 1.016$$

.

.

$$d_{72} = |e_{72} - \tilde{e}_2| = |0.317 - (-0.2341)| = 2.658$$

จากนั้นคำนวณค่าเฉลี่ยของทั้งสองกลุ่ม

$$\bar{d}_1 = \frac{41.55}{8} = 5.193 \text{ และ } \bar{d}_2 = \frac{27.65}{7} = 3.950$$

จากนั้นคำนวณค่ากำลังสองของความแตกต่างระหว่างค่า *d* กับค่าเฉลี่ยกำลังสองของแต่ละกลุ่ม

$$(d_{11} - \bar{d}_1)^2 = (2.0135 - 5.1936)^2 = 10.1140$$

$$(d_{12} - \bar{d}_2)^2 = (1.016 - 3.950)^2 = 8.6084$$

.

.

$$(d_{72} - \bar{d}_2)^2 = (2.658 - 3.950)^2 = 1.6693$$

และ $s = \sqrt{\frac{47.37053 + 56.26006}{15-2}} = 2.8234$

ดังนั้นค่าสถิติคือ

$$t_L^* = \frac{5.1936 - 3.950}{2.8234\sqrt{\frac{1}{8}+\frac{1}{7}}} = 0.851$$

เนื่องจาก t_L^* (0.851) < $t_{\text{ตาราง}}$ (2.160) ดังนั้นสรุปว่าความแปรปรวนของส่วนเหลือมีค่าคงที่

3.3 การทดสอบความเหมาะสมของสมการถดถอย

เนื่องจากสมการถดถอยเชิงเส้นอย่างง่ายนี้ใช้กับข้อมูลที่มีความสัมพันธ์เชิงเส้นตรงกัน การทดสอบความเหมาะสมของสมการถดถอย (lack of fit test) คือการทดสอบว่าสมการถดถอยที่ได้มีความเหมาะสมกับข้อมูลนั้นหรือไม่ หากตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้นตรงกัน สมการถดถอยที่ได้ไม่เหมาะที่จะใช้กับข้อมูลชุดนั้น การทดสอบความเหมาะสมมีข้อตกลงของตัวแปรตามดังนี้คือ (1) มีความเป็นอิสระต่อกัน (2) มีการแจกแจงแบบปกติและ (3) มีความแปรปรวนที่เท่ากัน นอกจากนี้การทดสอบนี้ที่ระดับของ X หนึ่งๆ ต้องมีค่ามากกว่า 1 ค่านั่นคือมีการวัดหรือจัดเก็บค่าซ้ำนั่นเอง เช่น การศึกษาความสัมพันธ์ระหว่างเชาวน์ปัญญากับผลสัมฤทธิ์ของการเรียนนั้นควรให้มีนักเรียนที่มีเชาวน์ปัญญาเท่ากันอยู่ในข้อมูลชุดนี้ เป็นต้น แต่หากเป็นการทดลองที่ควบคุมระดับของตัวแปรอิสระได้นั้นทำการทดลองซ้ำที่ระดับ X หนึ่งๆ แต่ไม่ใช่เป็นการวัดซ้ำหรืออ่านค่าซ้ำ การที่ต้องเก็บข้อมูล ณ ตำแหน่งเดิมหลายค่าเพื่อใช้ในการประมาณค่าความแปรปรวนของความคลาดเคลื่อน (σ^2) แต่ไม่จำเป็นว่าที่ทุกระดับของ X จะต้องมีค่า Y หลายค่า บางระดับสามารถมีค่า Y เพียงค่าเดียวได้ ในการทดสอบนี้มีสมมติฐานคือ H_0: $\beta_1 = 0$ และ H_1: $\beta_1 \neq 0$

สมมติให้ที่ระดับของ X_i ($i = 1,\ldots,m$) มีค่า Y_{ij} ($j = 1,\ldots,n_i$) อยู่ n_i ตัวโดย Y_{ij} คือค่า Y ตัวที่ j ของ X ในระดับที่ i ดังนั้นจำนวนข้อมูลทั้งหมดเท่ากับ n โดย $n = \sum_{i=1}^{m} n_i$ การทดสอบนี้แบ่งผลรวมกำลังสองความคลาดเคลื่อน (SSE) ออกเป็นสองส่วนคือ ผลรวมกำลังสองความคลาดเคลื่อนแท้จริง (pure error sum of squares หรือ $SSPE$) และผลรวมกำลังสองความคลาดเคลื่อนที่ไม่เหมาะสม (lack of fit sum of squares หรือ $SSLOF$) ดังนี้

$$SSE = SSPE + SSLOF \tag{3.8}$$

โดยที่ $SSPE$ เป็นผลรวมกำลังสองของการวัดความแปรผันของค่า Y ที่ตำแหน่ง X หนึ่งเมื่อเทียบกับค่าเฉลี่ยของ Y ณ ตำแหน่ง X นั้นดังนี้

$$SSPE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 \tag{3.9}$$

โดยองศาเสรีของ $SSPE$ เท่ากับ $\sum_{i=1}^{m}(n_i - 1) = n - m$

สำหรับ *SSLOF* เป็นผลรวมกำลังสองของการวัดความแปรผันของค่าเฉลี่ยของ Y ในแต่ละระดับของ X เมื่อเทียบกับค่าพยากรณ์ดังนี้

$$SSLOF = \sum_{i=1}^{m} n_i(\bar{Y}_i - \hat{Y}_i)^2 \tag{3.10}$$

โดยองศาเสรีของ *SSLOF* เท่ากับ n - 2

หากตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นตรงกันแล้วค่าเฉลี่ยของแต่ละกลุ่มควรมีค่าไม่แตกต่างจากค่าพยากรณ์มากแต่หากตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้นตรงกันแล้วค่าเฉลี่ยของแต่ละกลุ่มจะแตกต่างจากค่าพยากรณ์มาก สำหรับสถิติที่ใช้ในการทดสอบคือ F โดยมีสูตรการคำนวณดังนี้

$$F = \frac{SSLOF/(m-2)}{SSPE/(n-m)} = \frac{MSLOF}{MSPE} \tag{3.11}$$

หาก $F \geq F\alpha_{,\,m\,-\,2,\,n\,-\,m}$ แล้วสรุปว่าตัวแบบเชิงเส้นตรงที่ได้ไม่เหมาะสมกับข้อมูล

ตารางข้างล่างแสดงการวิเคราะห์ความแปรปรวนโดยแยกออกเป็นความคลาดเคลื่อนแท้จริงและความคลาดเคลื่อนที่ไม่เหมาะสม

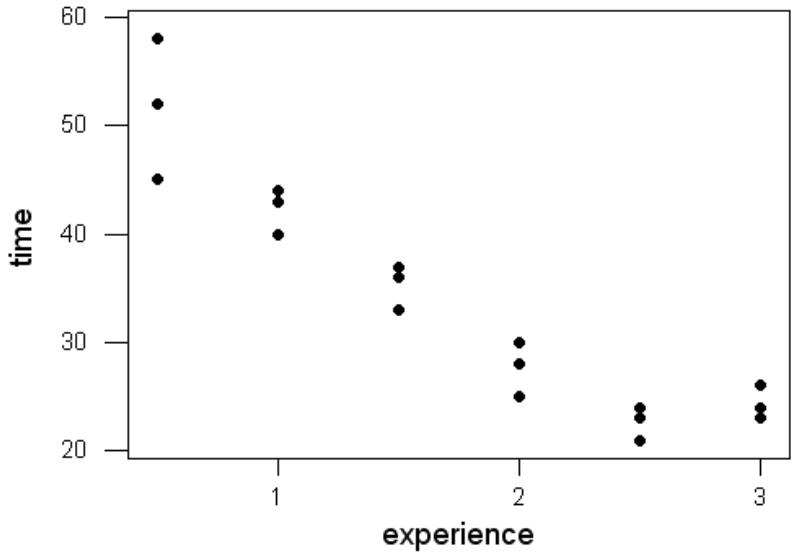
Source of variation	***SS***	***df***	***MS***	***F***
Regression	*SSR*	1	*MSR*	
Residual	*SSE*	$n-2$	*MSE*	
Lack of fit	*SSLOF*	$c-2$	*MSLOF*	*F*
Pure Error	*SSPE*	$n-c$	*MSPE*	
Total	*SST*	$n-1$		

ตัวอย่างที่ 3.3 ฝ่ายบุคคลต้องการศึกษาความสัมพันธ์ระหว่างอายุงานกับความเร็วที่ใช้ในการประกอบชิ้นส่วนรถยนต์ของพนักงานโดยมีข้อมูลดังตารางข้างล่าง จงทดสอบความเหมาะสมของสมการถดถอยที่ระดับนัยสำคัญ 0.05

อายุงาน (ปี)	เวลาที่ใช้ (นาที)
0.5	58
0.5	45
0.5	52
1.0	43
1.0	40
1.0	44
1.5	36
1.5	33
1.5	37
2.0	30
2.0	28
2.0	25
2.5	23
2.5	24
2.5	21
3.0	24
3.0	23
3.0	26

วิธีทำ

หากพิจารณาความสัมพันธ์ระหว่างข้อมูลจากกราฟที่ได้จาก MINITAB พบว่าข้อมูลมีความสัมพันธ์ที่ไม่เป็นเส้นตรงดังภาพข้างล่าง



จากข้อมูลข้างต้นจะได้สมการถดถอยคือ $\hat{Y}_i = 54.3 - 11.6X_i$ และมีค่า $SSE = 288.10$ ค่าเฉลี่ยของ Y ในแต่ละระดับของ X มีดังนี้ $\bar{y}_1 = 51.67,\ \bar{y}_2 = 42.33,\ \bar{y}_3 = 35.33,\ \bar{y}_4 = 27.67,\ \bar{y}_5 = 22.67$ และ $\bar{y}_6 = 24.33$ ดังนั้นสามารถคำนวณค่า $SSPE$ ได้ดังนี้

$$SSPE = (58 - 51.67)^2 + (45 - 51.67)^2 + \ldots + (23 - 24.33)^2 + (26 - 24.33)^2$$

$$= 124.00$$

ดังนั้น $SSLOF = 288.10 - 124.00 = 164.10$

องศาเสรีของ $SSPE = 18 - 6 = 12$

องศาเสรีของ $SSLOF = 6 - 2 = 4$

และค่าสถิติ F คือ

$$F = \frac{SSLOF/(m-2)}{SSPE/(n-m)}$$

$$= \frac{164.10/4}{124.00/12} = 3.97$$

หากเปรียบเทียบค่าสถิติที่ได้กับค่าวิกฤตจากตาราง $F_{0.05,\,4,\,12} = 3.26$ พบว่า

$$F = 3.97 > F_{0.05,\,4,\,12} = 3.26$$

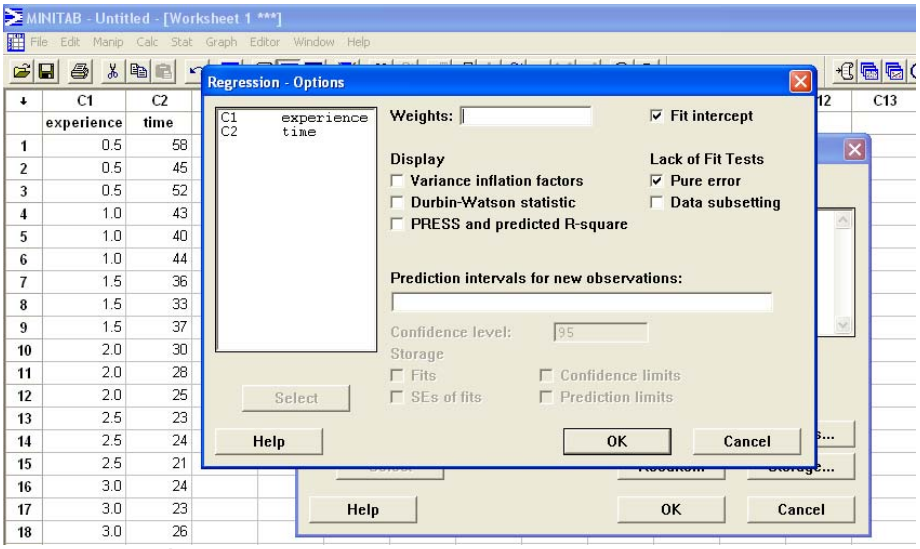
ดังนั้นสรุปว่าสมการถดถอยที่ได้นี้ไม่เหมาะสมกับข้อมูลชุดนี้เนื่องจากข้อมูลไม่เป็นเส้นตรง

หมายเหตุ

1. หากการทดสอบความสัมพันธ์เชิงเส้นตรงของสมการถดถอยแสดงให้เห็นว่าไม่มีความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรทั้งสองหรือตัวแปรอิสระไม่สามารถพยากรณ์

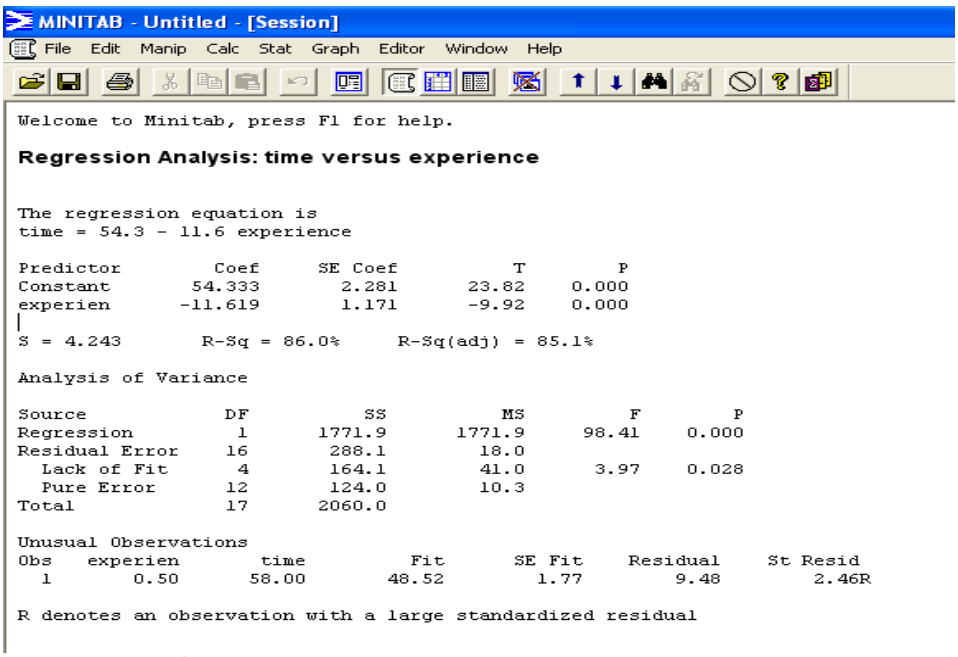
ตัวแปรตามได้ บางครั้งหากทำการทดสอบความเหมาะสมของสมการถดถอยแล้วจะพบว่าสมการถดถอยเชิงเส้นอย่างง่ายไม่เหมาะสมกับข้อมูลชุดนี้ซึ่งที่จริงแล้วข้อมูลชุดนี้มีความสัมพันธ์กันแต่ไม่เป็นเส้นตรงก็ได้ ดังนั้นการตรวจสอบความเหมาะสมของสมการถดถอยจึงเป็นสิ่งสำคัญที่ควรตรวจสอบ

2. นักวิจัยควรระวังในการเก็บข้อมูลซ้ำต้องพยายามให้ข้อมูลที่เก็บในระดับเดียวกันมีความเป็นอิสระต่อกัน
3. หากไม่สามารถเก็บข้อมูลให้มีค่า x ที่ระดับเดียวกันได้ก็สามารถทดสอบความเหมาะสมได้ โดยนำข้อมูลที่มีค่า x ใกล้กันมาจัดเป็นกลุ่มเดียวกัน
4. ผู้อ่านสามารถใช้ MINITAB ช่วยในการทดสอบความเหมาะสมได้โดยการเข้าไปที่ “Option” ของการวิเคราะห์สมการถดถอยแล้วเลือก “Pure error” ภายใต้ “Lack of Fit Tests” ดังภาพที่ 3.13



ภาพที่ 3.13 หน้าจอการทดสอบความเหมาะสม

ผลลัพธ์ที่ได้จากการใช้ MINITAB แสดงดังภาพที่ 3.14 ในรูปของตาราง ANOVA โดยในส่วนของการวิเคราะห์ความเหมาะสมจะดูจากค่า F และ p-value ของ Lack of Fit จะพบว่าสอดคล้องกับผลลัพธ์ในตัวอย่างที่ 3.3



ภาพที่ 3.14 หน้าจอผลลัพธ์การทดสอบความเหมาะสม

3.4 การแก้ไขข้อมูลที่ไม่เป็นไปตามข้อตกลง

การวิเคราะห์ข้อมูลที่ไม่เป็นไปตามข้อตกลงสามารถทำได้สองวิธีคือ

(1) หารูปแบบการถดถอยที่เหมาะสมในการวิเคราะห์ ข้อดีของวิธีนี้คือนักวิเคราะห์สามารถเข้าใจและตีความสิ่งที่ได้ไม่ยากเนื่องจากข้อมูลยังคงเดิมแต่ข้อเสียคือวิธีนี้อาจยุ่งยากและต้องใช้ความรู้ทางสถิติที่ค่อนข้างสูงและบางครั้งอาจต้องการข้อมูลจำนวนมากในการวิเคราะห์

(2) แปลงข้อมูล (transform) ให้เป็นไปตามข้อตกลงแล้วทำการวิเคราะห์โดยใช้ตัวแบบเดิม วิธีนี้ง่ายกว่าวิธีแรกเนื่องจากในปัจจุบันโปรแกรมสำเร็จรูปทางสถิติมีคำสั่งที่ช่วยในการแปลงข้อมูลและสามารถใช้กับข้อมูลที่มีขนาดน้อยได้ แต่ข้อเสียคือ ยากแก่การตีความพารามิเตอร์ที่ได้จากการแปลงข้อมูล

การแก้ไขข้อมูลที่ไม่เป็นไปตามข้อตกตงสามารถทำได้หลายวิธีดังนี้

3.4.1 การแก้ไขความสัมพันธ์ที่ไม่เป็นเส้นตรง

บ่อยครั้งอาจพบว่าหากข้อมูลไม่มีความสัมพันธ์เชิงเส้นตรงแล้วข้อมูลอาจมีความสัมพันธ์แบบโพลิโนเมียลหรือเอ็กซ์โปเนนเชียล เช่น $E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X^2$ หรือ $E(y) = \beta_0 \beta_1^X$ เป็นต้น

3.4.2 การแก้ไขความแปรปรวนที่ไม่คงที่ของความคลาดเคลื่อน

หากความแปรปรวนของความคลาดเคลื่อนไม่คงที่แบบเป็นระบบสามารถแก้ไขโดยการใช้วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (weighted least square method) ในการประมาณค่าพารามิเตอร์

3.4.3 การแก้ไขความไม่เป็นอิสระของความคลาดเคลื่อน

หากความคลาดเคลื่อนมีความสัมพันธ์ต่อกันสามารถแก้ไขโดยการใช้ตัวแบบที่ยอมให้ความคลาดเคลื่อนมีความสัมพันธ์กัน

3.4.4 การแก้ไขความคลาดเคลื่อนที่ไม่มีการแจกแจงแบบปกติ

ส่วนมากแล้วการที่ความคลาดเคลื่อนไม่มีการแจกแจงแบบปกติและมีความแปรปรวนไม่คงที่มักเกิดขึ้นพร้อมๆ กัน การแปลงข้อมูลนอกจากจะช่วยให้ความคลาดเคลื่อนมีความแปรปรวนคงที่แล้วยังช่วยให้ความคลาดเคลื่อนมีการแจกแจงปกติอีกด้วย ดังนั้นจึงควรแปลงข้อมูลให้มีความแปรปรวนที่คงที่ก่อนแล้วตรวจสอบว่าความคลาดเคลื่อนมีการแจกแจงปกติหรือไม่หลังจากการแปลงข้อมูลแล้ว

3.4.5 การแก้ไขข้อมูลในกรณีที่ไม่ได้รวมตัวแปรที่สำคัญเข้าในตัวแบบ

หากการวิเคราะห์ความคลาดเคลื่อนชี้ให้เห็นว่ามีการละตัวแปรที่สำคัญไปควรจะเพิ่มตัวแปรเหล่านั้นเข้าไปในตัวแบบ

3.5 การแปลงข้อมูล

การแปลงข้อมูลสามารถทำได้ทั้งการแปลงตัวแปรอิสระและตัวแปรตามอย่างใดอย่างหนึ่งหรือแปลงทั้งตัวแปรอิสระและตัวแปรตามพร้อมกันก็ได้ ในที่นี้จะกล่าวถึงการแปลงข้อมูลให้มีความสัมพันธ์เป็นเส้นตรงและการแปลงข้อมูลเพื่อให้ความแปรปรวนมีค่าคงที่

3.5.1 การแปลงข้อมูลที่ไม่มีความสัมพันธ์เป็นเส้นตรง

เนื่องจากข้อตกลงหนึ่งของการใช้สมการถดถอยเชิงเส้นอย่างง่ายคือ ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์เชิงเส้นกัน การวิเคราะห์ว่าตัวแปรมีความสัมพันธ์เชิงเส้นหรือไม่สามารถทดสอบโดยการใช้การทดสอบความเหมาะสมของสมการถดถอยดังที่ได้กล่าวไว้แล้วหรือการพิจารณาจากแผนภาพกระจายระหว่างตัวแปรทั้งสอง

หากความคลาดเคลื่อนไม่มีการแจกแจงแบบปกติแต่ใกล้เคียงการแจกแจงแบบปกติและมีความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่แล้วควรทำการแปลงตัวแปรอิสระทั้งนี้เนื่องจากการแปลงตัวแปรตามอาจทำให้ความคลาดเคลื่อนมีการแจกแจงที่ไม่ห่างไกลจากการแจกแจงปกติมากขึ้นหรือความแปรปรวนของความคลาดเคลื่อนที่แตกต่างกันมากขึ้นก็ได้

Montgomery และ Peck (1992, p. 89-90) ได้เสนอรูปแบบในการแปลงตัวแปรอิสระและ/หรือตัวแปรตามโดยดูจากแผนภาพกระจายระหว่างตัวแปรอิสระและตัวแปรตาม หากพิจารณาฟังก์ชันเอ็กซ์โปเนนเชียล

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

จะพบว่าสามารถแปลงฟังก์ชันนี้ให้เป็นเส้นตรงได้โดยการเปลี่ยนให้อยู่ในรูปของลอการิทึมดังนี้

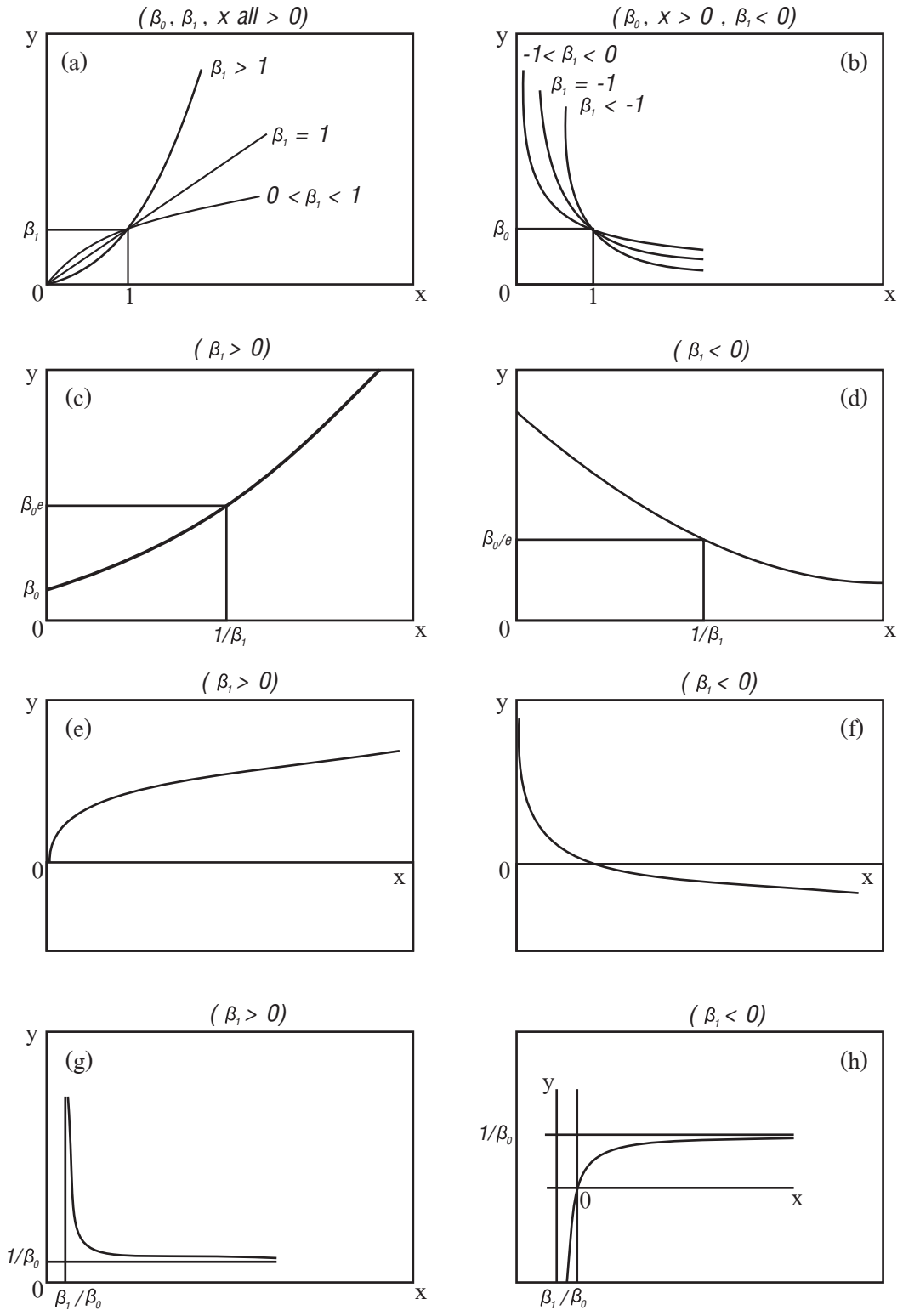
$$\ln Y = \ln \beta_0 + \beta_1 X + \ln \varepsilon$$

หรือ

$$Y' = \beta_0' + \beta_1 X + \varepsilon'$$

โดยที่ต้องตรวจสอบว่า $\varepsilon' = \ln \varepsilon$ นั้นมีการแจกแจงแบบปกติหรือไม่หลังจากการสร้างสมการถดถอยของข้อมูลที่ได้จากการแปลงแล้ว

รูปแบบการแปลงที่เหมาะสมแบบต่างๆ ตามฟังก์ชันของข้อมูลแสดงดังภาพที่ 3.15 และตารางที่ 3.1



ภาพที่ 3.15 ฟังก์ชันของข้อมูลแบบต่างๆ

ที่มา : Montgomery & Peck, 1992, p.90

ตารางที่ 3.1 ฟังก์ชันการแปลงข้อมูลแบบต่างๆ

รูป	ฟังก์ชันของข้อมูล	รูปแบบการแปลง	ฟังก์ชันหลังการแปลง
a และ b	$Y = \beta_0 X^{\beta_1}$	$Y' = \log Y,\ X' = \log X$	$Y' = \log \beta_0 + \beta_1 X'$
c และ d	$Y = \beta_0 e^{\beta_1 X}$	$Y' = \ln Y$	$Y' = \ln \beta_0 + \beta_1 X$
e และ f	$Y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y' = \beta_0 + \beta_1 X'$
g และ h	$Y = \frac{X}{\beta_0 X - \beta_1}$	$Y' = \frac{1}{Y},\ X' = \frac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$

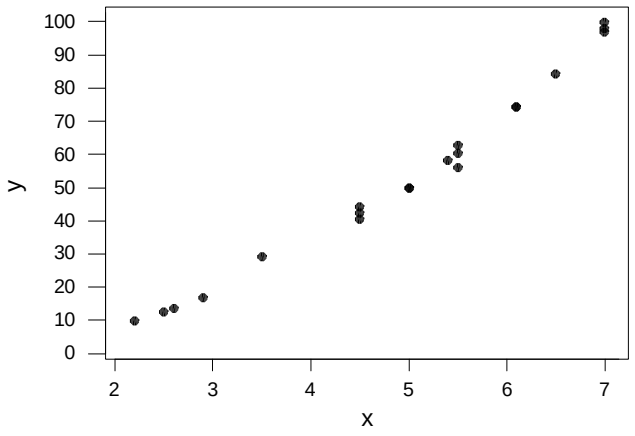
ตัวอย่างที่ 3.4 จากข้อมูลข้างล่าง

X	5.5	5.5	2.9	4.5	5.0	5.5	6.1	7.0	7.0	2.2
Y	62.7	56.2	16.8	40.5	50.0	50.0	74.4	100.0	97.0	9.7
X	7.0	6.5	2.6	5.5	4.5	4.5	5.4	2.5	6.1	3.5
Y	98.0	84.5	13.5	60.5	42.3	44.2	58.3	12.5	74.4	29.0

จงทดสอบความเหมาะสมของตัวแบบที่ระดับนัยสำคัญ 0.05 และหากพบว่าตัวแบบเชิงเส้นไม่เหมาะสมจงแปลงข้อมูล

วิธีทำ

จากแผนภาพกระจายระหว่างตัวแปรทั้งสองพบว่าข้อมูลไม่มีความสัมพันธ์เชิงเส้นตรงกัน ดังภาพข้างล่าง



จากข้อมูลข้างต้นจะได้สมการถดถอยคือ $\hat{Y}_i = -37.3 + 18.4X_i$ และเมื่อทดสอบความเหมาะสมของตัวแบบพบว่าตัวแบบเชิงเส้นไม่เหมาะสมเนื่องจาก p-value เท่ากับ 0.003 ถึงแม้ว่าตัวแบบเชิงเส้นจะไม่เหมาะสมแต่พบว่าค่า R-Sq(adj) อยู่ในระดับที่สูงคือ 97.4%

Regression Analysis: y versus x

The regression equation is
y = - 37.3 + 18.4 x

Predictor	Coef	SE Coef	T	P
Constant	-37.292	3.578	-10.42	0.000
x	18.4245	0.6932	26.58	0.000

S = 4.638 R-Sq = 97.5% R-Sq(adj) = 97.4%

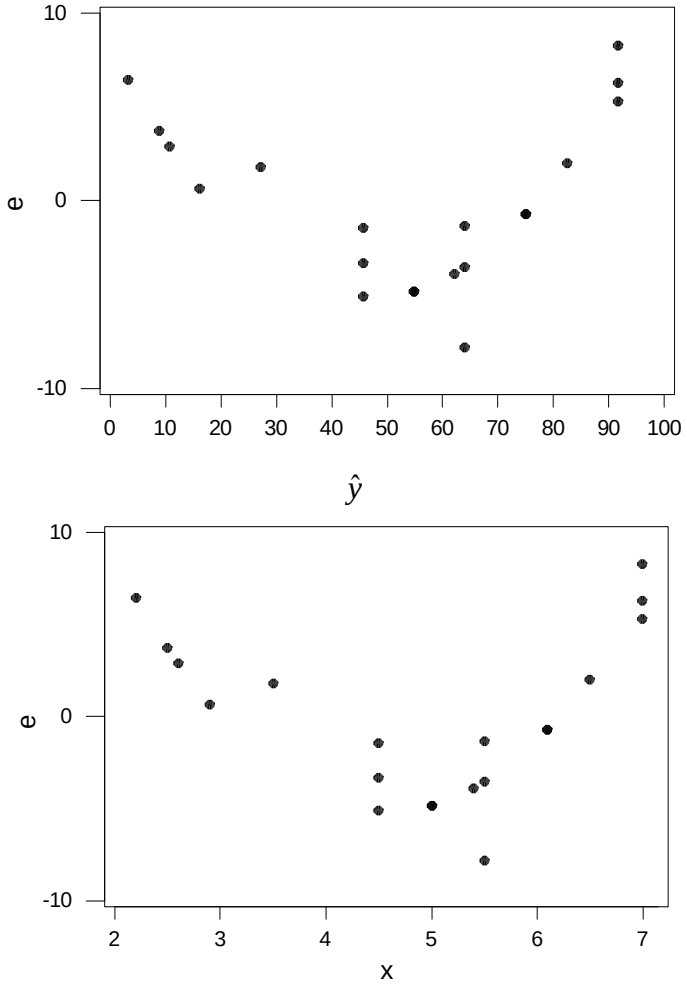
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15197	15197	706.44	0.000
Residual Error	18	387	22		
Lack of Fit	10	354	35	8.48	0.003
Pure Error	8	33	4		
Total	19	15584			

7 rows with no replicates

Lack of fit test
Possible curvature in variable x (P-Value = 0.000)
Possible lack of fit at outer X-values (P-Value = 0.016)
Overall lack of fit test is significant at P = 0.000

นอกจากนี้หากพิจารณาแผนภาพกระจายระหว่างค่าพยากรณ์กับส่วนเหลือพบว่าไม่มีความสัมพันธ์เชิงเส้นกันเช่นเดียวกับแผนภาพกระจายระหว่างตัวแปรอิสระกับส่วนเหลือดังภาพข้างล่าง



เมื่อพิจารณาแผนภาพกระจายระหว่างตัวแปรอิสระกับตัวแปรตามพบว่าตัวแปรทั้งสองมีความสัมพันธ์แบบเอ็กซ์โปเนนเชียลซึ่งตรงกับรูปแบบในภาพที่ 3.15 a จึงควรแปลงข้อมูลของตัวแปรทั้งสองเป็น $\log X$ และ $\log Y$ ดังนี้

$$\log Y = \log \beta_0 + \beta_1 \log X + \log \varepsilon$$

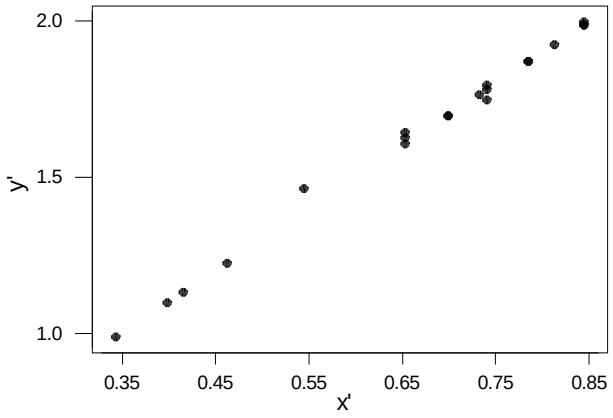
หรือ

$$Y' = \beta_0' + \beta_1 X' + \varepsilon'$$

ดังนั้นข้อมูลของตัวแปรทั้งสองหลังจากการแปลงด้วยลอการิทึมมีดังนี้คือ

X'	0.7404	0.7404	0.4624	0.6532	0.6990	0.6990	0.7853	0.8451	0.8451	0.3424
Y'	1.7973	1.7497	1.2253	1.6075	1.6990	1.6990	1.8716	2.0000	1.9868	0.9868
X'	0.8451	0.8129	0.4150	0.7404	0.6532	0.6532	0.7324	0.3979	0.7853	0.5441
Y'	1.9912	1.9269	1.1303	1.7818	1.6263	1.6454	1.7657	1.0969	1.8716	1.4624

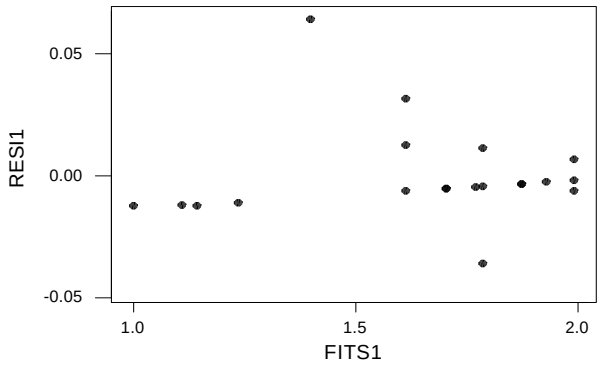
เมื่อนำข้อมูลที่แปลงมาสร้างแผนภาพกระจายระหว่างตัวแปรทั้งสองพบว่ามีความสัมพันธ์เชิงเส้นดังภาพข้างล่าง



เมื่อนำข้อมูลที่แปลงแล้วมาสร้างสมการถดถอยได้ $\hat{Y}_i' = 0.322 + 1.98X_i'$ และพบว่า R-Sq(adj) มีค่าสูงขึ้นเป็น 99.6% นอกจากนี้เมื่อทดสอบความเหมาะสมของตัวแบบพบว่าตัวแบบมีความเหมาะสมด้วย p-value เท่ากับ 0.130

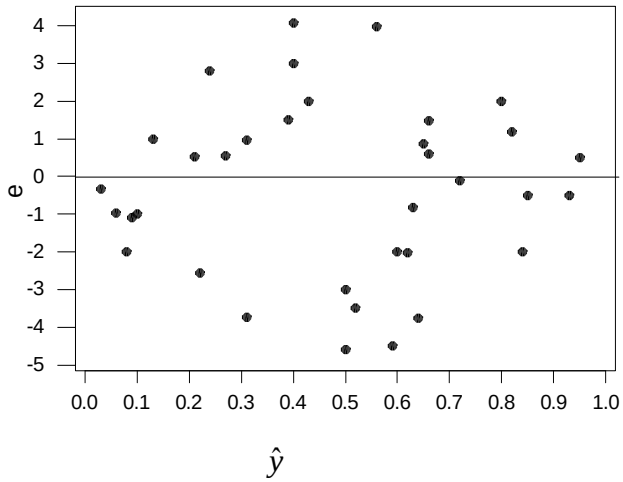
```
The regression equation is
Y′ = 0.322 + 1.98 x′

Predictor        Coef     SE Coef          T        P
Constant      0.32213     0.02068      15.58    0.000
X′            1.97725     0.03011      65.67    0.000

S = 0.02054     R-Sq = 99.6%     R-Sq(adj) = 99.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1.8202      1.8202   4312.82    0.000
Residual Error    18      0.0076      0.0004
 Lack of Fit      10      0.0056      0.0006      2.26    0.130
 Pure Error        8      0.0020      0.0002
Total             19      1.8278

7 rows with no replicates

Unusual Observations
Obs      x′      y′       Fit      SE Fit   Residual    St Resid
 20    0.544  1.46240   1.39788   0.00595   0.06451     3.28R

R denotes an observation with a large standardized residual

No evidence of lack of fit (P > 0.1)
```

ผลลัพธ์ที่ได้สอดคล้องกับแผนภาพกระจายระหว่างค่าพยากรณ์กับส่วนเหลือที่ไม่แสดงรูปแบบใดและมีการกระจายตัวอย่างสุ่มดังภาพข้างล่าง ดังนั้นแสดงให้เห็นว่าการแปลงข้อมูลนั้นทำให้ข้อมูลมีความเหมาะสมกับตัวแบบสมการเชิงเส้นอย่างง่าย



3.5.2 การแปลงข้อมูลที่ไม่มีการแจกแจงแบบปกติและมีความแปรปรวนไม่คงที่

โดยทั่วไปหากพบว่าข้อมูลไม่มีการแจกแจงแบบปกติแล้วมักจะพบว่าข้อมูลมีความแปรปรวนที่ไม่คงที่ ในการสร้างสมการถดถอยนั้นตัวแปรตามควรมีการแจกแจงแบบปกติ ดังนั้นการแปลงค่าตัวแปรตามจึงมีความจำเป็นและในขณะเดียวกันการแปลงข้อมูลของตัวแปรตามนั้นยังอาจช่วยให้ความสัมพันธ์ของตัวแปรทั้งสองเป็นเส้นตรงอีกด้วย สาเหตุหลักของความแปรปรวนที่ไม่คงที่นั้นส่วนใหญ่มาจากการที่ความแปรปรวน (σ^2) ของตัวแปรตามมีความสัมพันธ์กับค่าเฉลี่ย ($E(Y)$) เช่น หาก Y มีการแจงแจงแบบป๊วซองแล้วความแปรปรวนของ Y จะมีค่าเท่ากับค่าเฉลี่ย ดังนั้นการแก้ไขความแปรปรวนที่ไม่คงที่นั้นสามารถทำได้โดยการถอดรากของ Y หรือ $Y' = \sqrt{Y}$ หรือหาก Y เป็นข้อมูลสัดส่วนซึ่งมีค่าอยู่ในช่วง 0 กับ 1 แล้วแผนภาพระหว่างค่าพยากรณ์กับส่วนเหลือจะมีลักษณะดังภาพที่ 3.16 ในกรณีเช่นนี้ควรแปลง Y ด้วย arcsin หรือ $Y' = \sin^{-1}(\sqrt{Y})$ เป็นต้น



ภาพที่ 3.16 ความสัมพันธ์ระหว่างค่าพยากรณ์และส่วนเหลือ

Montgomery & Peck (1992, p. 98) เสนอรูปแบบการแปลงตัวแปรตามไว้ดังตารางที่ 3.2

ตารางที่ 3.2 รูปแบบการแปลงโดยพิจารณาฟังก์ชันของความแปรปรวน

ความสัมพันธ์ระหว่าง σ^2 กับ $E(Y)$	**รูปแบบการแปลง**
$\sigma^2 \ \alpha$ ค่าคงที่	$Y' = Y$ (ไม่ต้องแปลงข้อมูล)
$\sigma^2 \ \alpha \ E(Y)$	$Y' = \sqrt{Y}$ (กรณี y แจกแจงแบบปัวซอง)
$\sigma^2 \ \alpha \ E(Y)[1 - E(Y)]$	$Y' = \sin^{-1}(\sqrt{Y})$ (กรณี $0 \leq Y \leq 1$)
$\sigma^2 \ \alpha \ [E(Y)]^2$	$Y' = \ln(Y)$
$\sigma^2 \ \alpha \ [E(Y)]^3$	$Y' = Y^{-\frac{1}{2}}$
$\sigma^2 \ \alpha \ [E(Y)]^4$	$Y' = Y^{-1}$

จากรูปแบบการแปลงตัวแปรข้างต้นพบว่าการแปลงโดยใช้ลอการิทึมนั้นเหมาะกับค่า Y ที่มีค่าความแปรปรวนเพิ่มขึ้นหรือลดลงแบบเอ็กซ์โปเนนเชียลและเมื่อค่าเฉลี่ยของ Y มีกำลังเพิ่มมากขึ้นการแปลงจะใช้ฟังก์ชันที่ต่ำลงเพื่อลดความแปรปรวนของ Y ที่เพิ่มขึ้นอย่างรวดเร็ว เช่น การใช้ส่วนกลับของ Y เป็นต้น

การที่ค่าของตัวแปรตามไม่คงที่จะส่งผลให้ความแปรปรวนของส่วนเหลือไม่คงที่ด้วย การตรวจสอบว่าค่าของตัวแปรตามมีความแปรปรวนคงที่หรือไม่อย่างง่ายคือการวาดแผนภาพกระจายระหว่างของส่วนเหลือกับค่าพยากรณ์หรือกับค่าของตัวแปรตาม โดยแผนภาพจะช่วยให้การเลือกรูปแบบการแปลงข้อมูลทำได้ง่ายขึ้น การตรวจสอบและปรับแก้ความแปรปรวนให้คงที่

นั้นจำเป็นต้องทำหากใช้วิธีกำลังสองน้อยที่สุดในการสร้างสมการถดถอยเนื่องจากค่าประมาณของพารามิเตอร์ที่ได้จะไม่เป็นตัวประมาณค่าที่มีความแปรปรวนน้อยที่สุดแต่ยังคงเป็นตัวประมาณค่าที่ไม่เอนเอียง การที่ตัวประมาณค่าไม่มีความแปรปรวนที่น้อยที่สุดนั้น ทำให้ค่าสัมประสิทธิ์ของสมการถดถอยมีค่าส่วนเบี่ยงเบนมาตรฐานที่ใหญ่ขึ้นส่งผลให้ความถูกต้องของการประมาณค่าและการพยากรณ์ลดลง ซึ่งอาจส่งผลให้การทดสอบสมมติฐานของค่าสัมประสิทธิ์ไม่ถูกต้องไปด้วย

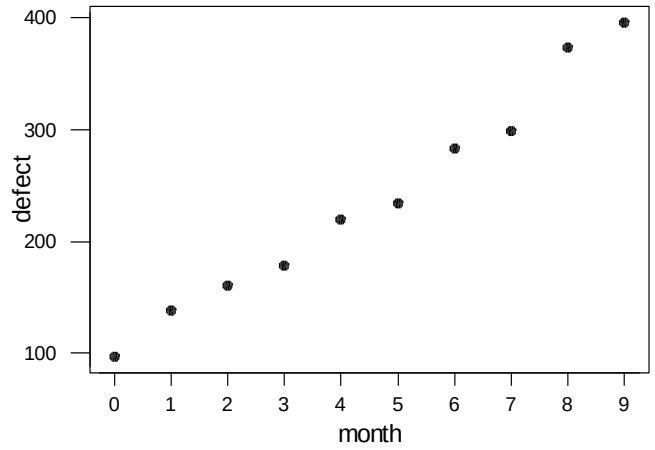
หลังจากการแปลงข้อมูลแล้วหากทำการพยากรณ์ค่า Y ต้องมีการแปลงค่าพยากรณ์กลับมาสู่รูปแบบเดิมหรือหน่วยเดิม เช่น หากแปลงค่า Y โดยการถอดรากต้องยกกำลังสองของค่าพยากรณ์ ($\hat{Y}^2$) เพื่อให้ค่าพยากรณ์อยู่ในหน่วยเดียวกันกับข้อมูลเดิม เป็นต้น

ตัวอย่าง 3.5 โรงงานผลิตแก้วต้องการศึกษาจำนวนแก้วที่มีตำหนิต่อเดือน (หน่วยเป็นพันชิ้น) เป็นจำนวน 10 เดือน โดยมีข้อมูลดังนี้

X (เดือนที่)	0	1	2	3	4	5	6	7	8	9
Y (จำนวน)	97	138	161	178	220	234	283	299	374	396

วิธีทำ

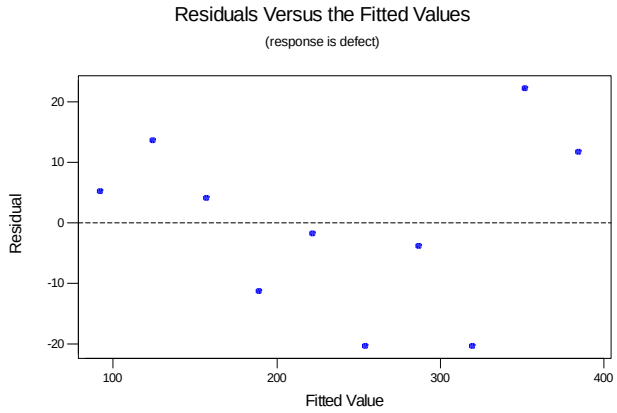
จากข้อมูลดังกล่าวเมื่อนำมาวาดแผนภาพกระจายระหว่างตัวแปรทั้งสองไม่พบความผิดปกติแต่อย่างใดดังภาพและเมื่อสร้างสมการถดถอยระหว่างตัวแปรทั้งสองจะได้สมการถดถอยคือ $\hat{Y}_i = 91.8 + 32.5X_i$



จากตารางการวิเคราะห์ความแปรปรวนที่ได้จาก MINITAB พบว่าค่า R^2 อยู่ในระดับที่สูงมากคือ 97.9%

```
The regression equation is
defect = 91.8 + 32.5 month

Predictor        Coef     SE Coef          T        P
Constant       91.764       8.893      10.32    0.000
month          32.497       1.666      19.51    0.000

S = 15.13       R-Sq = 97.9%     R-Sq(adj) = 97.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       87124       87124    380.53    0.000
Residual Error     8        1832         229
Total              9       88956
```

แต่เมื่อพิจารณาจากแผนภาพกระจายระหว่างค่าพยากรณ์กับส่วนเหลือพบว่าที่ค่าพยากรณ์ต่ำๆ นั้นส่วนเหลือจะมีค่าเป็นบวกในทำนองเดียวกันกับค่าพยากรณ์สูงๆ แต่เมื่อค่าพยากรณ์มีค่ากลางๆ นั้นส่วนเหลือจะมีค่าเป็นลบ



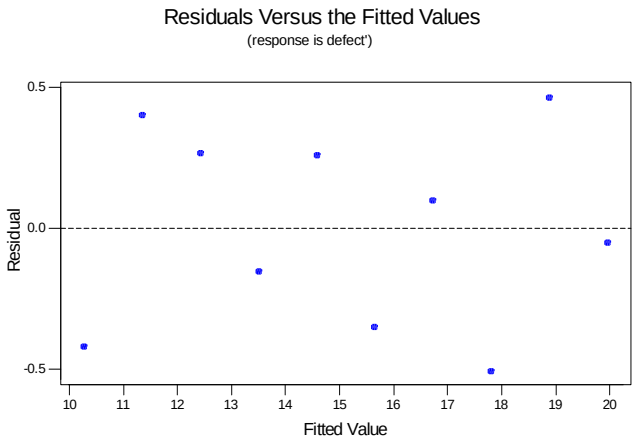
แสดงให้เห็นว่าความแปรปรวนของค่าคลาดเคลื่อนนั้นไม่คงที่ ในกรณีนี้จึงควรแปลงค่า Y โดยการถอดรากหรือ $Y' = \sqrt{Y}$ หลังจากการแปลงข้อมูลแล้วพบว่าได้สมการถดถอยคือ

$$\hat{Y}_i' = 10.3 + 1.08 X_i$$

จากตารางการวิเคราะห์ความแปรปรวนที่ได้จาก MINITAB พบว่าค่า R^2 มีค่าเพิ่มมากขึ้นเล็กน้อยเป็น 98.9%

```
The regression equation is
defect' = 10.3 + 1.08 month

Predictor        Coef     SE Coef          T        P
Constant      10.2694      0.2190      46.89    0.000
month         1.07590     0.04102      26.23    0.000

S = 0.3726      R-Sq = 98.9%     R-Sq(adj) = 98.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      95.498      95.498    687.89    0.000
Residual Error     8       1.111       0.139
Total              9      96.609
```

เมื่อพิจารณาแผนภาพกระจายระหว่างค่าพยากรณ์กับส่วนเหลือไม่พบรูปแบบที่ผิดปกติแต่อย่างใดดังภาพข้างล่างแสดงว่าตัวแบบที่ได้จากข้อมูลที่แปลงแล้วมีความเหมาะสม



3.5.3 การแปลงข้อมูลโดยวิธี Box-Cox transformation

เมื่อพบว่าตัวแปรตามมีการแจกแจงที่ไม่ปกติหรือมีความแปรปรวนที่ไม่คงที่ดังในตัวอย่างที่ผ่านมา การเลือกรูปแบบการแปลงโดยการพิจารณาจากแผนภาพอาจมีความยุ่งยาก Box และ Cox (1964) ได้คิดค้นสูตรการแปลงข้อมูลโดยจัดให้อยู่ในรูปของการยกกำลังของ Y หรือ Y^{λ} โดยที่ λ เป็นพารามิเตอร์ที่ใช้ในการกำหนดรูปแบบการแปลง ดังนี้

$\lambda = -1.0$ คือ $Y' = \frac{1}{Y}$

$\lambda = -0.5$ คือ $Y' = \frac{1}{\sqrt{Y}}$

$\lambda = 0.0$ คือ $Y' = \ln(Y)$

$\lambda = 0.5$ คือ $Y' = \sqrt{Y}$

$\lambda = 1.0$ คือ $Y' = Y$ (ไม่มีการแปลงข้อมูล)

$\lambda = 2.0$ คือ $Y' = Y^2$

แนวคิดของวิธีการนี้คือการสร้างสมการถดถอยโดยวิธีความน่าจะเป็นสูงสุด (maximum likelihood method) ในการประมาณค่า λ และทำการคัดเลือกซ้ำโดยพิจารณาจาก λ ที่ให้ผลรวมกำลังสองของความคลาดเคลื่อน (*SSE*) ที่ได้ หาก λใดให้ค่า *SSE* ที่ต่ำสุดนั้นเป็น λ ที่เหมาะสมที่สุด เนื่องจากการใช้ค่า λ ที่แตกต่างกันจะให้ค่า Y มีหน่วยที่แตกต่างกันทำให้ *SSE* ที่ได้จากการสมการถดถอยนั้นๆ ไม่สามารถนำมาเปรียบเทียบกันได้ดังนั้นจึงต้องแปลงค่า Y ในอยู่ในรูปเดียวกัน ($Y^{(\lambda)}$) ดังนี้

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda \dot{Y}^{\lambda - 1}}, & \lambda \neq 0 \\ \dot{Y} \ln(Y), & \lambda = 0 \end{cases}$$

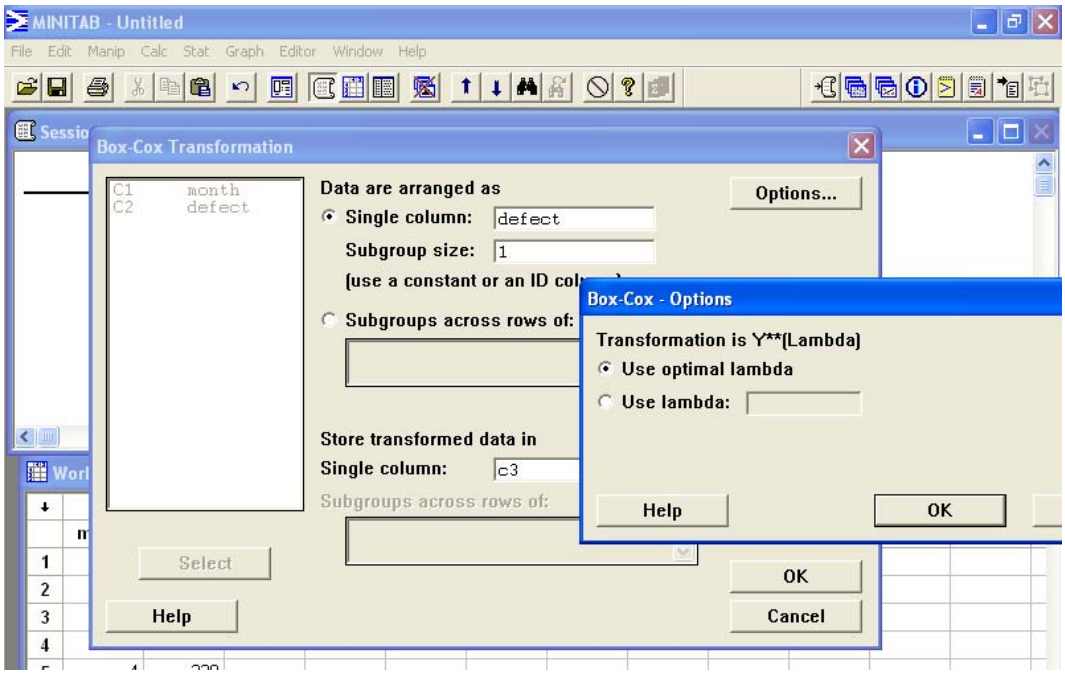
โดย $\dot{Y} = \ln^{-1}\left[\dfrac{\sum \ln(Y)}{n}\right]$

จากนั้นนำค่า $Y^{(\lambda)}$ ที่ได้มาสร้างสมการถดถอยกับตัวแปรอิสระเพื่อหาค่า *SSE* แล้วเปรียบเทียบเพื่อหาค่า λ ที่ให้ *SSE* ที่ต่ำที่สุด การคำนวณด้วยมือนั้นจะยุ่งยากและเสียเวลามาก ในปัจจุบันนี้มีโปรแกรมสำเร็จรูปที่ช่วยในการหาค่า λ ที่เหมาะสมพร้อมทั้งแปลงข้อมูลโดยการใช้ค่า λ นั้นเช่น โปรแกรม MINITAB เป็นต้น

การใช้โปรแกรม MINITAB มีขั้นตอนดังนี้

1. เลือก “Stat”
2. เลือก “Control Charts”
3. เลือก “Box-Cox Transformation..”
4. ระบุค่าตัวแปร Y ที่ต้องการแปลงค่าลง “Single column:”
5. ระบุขนาดของกลุ่มย่อย (subgroup) ลงในช่อง “Subgroup size” หากไม่มีกลุ่มย่อยให้ใส่ 1
6. หากต้องการเก็บค่าที่แปลงแล้วให้ระบุชื่อคอลัมน์ใน “Store transform data in single column:”

7. นักวิจัยสามารถเลือกที่จะระบุค่า λ หรือให้โปรแกรมหาค่า λ ที่เหมาะสมให้โดยคลิกที่ “Options…” จากนั้นระบุค่า λ ลงในช่อง “Use lambda:” แต่หากไม่ระบุค่าโปรแกรมจะหาค่า λ ที่เหมาะสมให้เอง
8. จากนั้นคลิก “OK” ดังภาพที่ 3.17

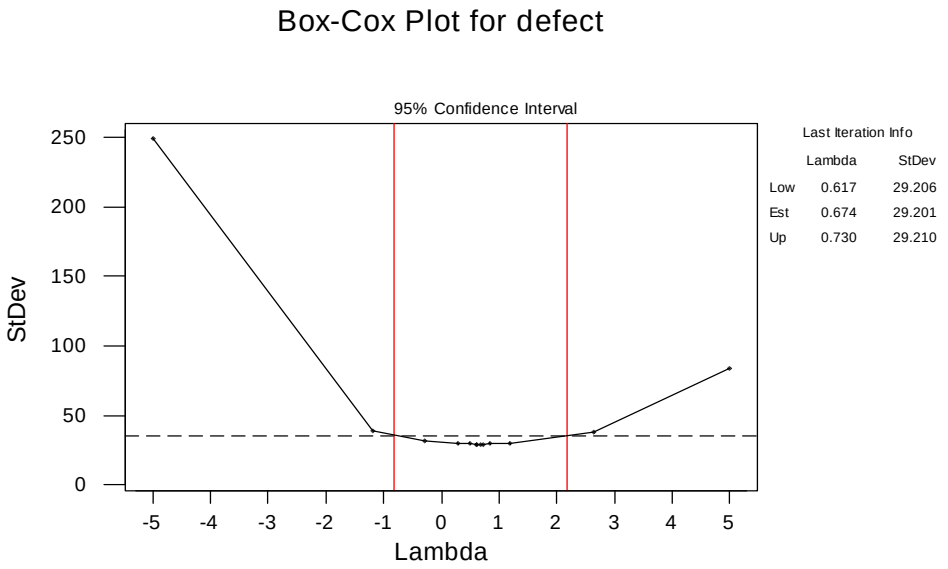


ภาพที่ 3.17 หน้าจอการหาค่า λ ที่เหมาะสม

ตัวอย่างที่ 3.6 จากข้อมูลในตัวอย่างที่ 3.5 หากใช้โปรแกรม MINITAB ในการหาค่า λ ที่เหมาะสม พร้อมทั้งคำนวณค่า Y ที่แปลงแล้ว

วิธีทำ

ข้อมูลชุดนี้ไม่มีกลุ่มย่อยดังนั้น subgroup จึงเท่ากับ 1 จากนั้นให้โปรแกรมหาค่า λ ที่เหมาะสมให้ หลังจากการคำนวณโปรแกรมจะแสดงกราฟดังภาพ



จากภาพพบว่าค่า λ ที่เหมาะสมคือ 0.674 และช่วงความเชื่อมั่นที่ 95% ของค่า λ คือ (0.617, 0.73) ซึ่งค่า λ มีค่าใกล้เคียงกับ 0.5 (ที่ $\lambda = 0.5$ นั้นการแปลงที่เหมาะสมคือ การถอดราก หรือ $Y' = \sqrt{Y}$) ดังในตัวอย่างที่ 3.5 ที่แปลงข้อมูลโดยใช้การถอดราก

หมายเหตุ

1. หลังจากการแปลงข้อมูลแล้วค่าประมาณของสัมประสิทธิ์ (b_0 และ b_1) ที่ได้นั้นจะมีคุณสมบัติของสมการถดถอยตามข้อมูลที่แปลงแล้วมิใช่ข้อมูลเดิม

2. เนื่องจากค่า λ เป็นค่าที่ได้จากข้อมูล หากมีค่าที่ใกล้เคียงกับค่า λ ที่สามารถจัดรูปแบบการแปลงได้ควรจะเลือกใช้รูปแบบการแปลงที่ใกล้เคียงเพื่อง่ายแก่การแปลงข้อมูลดังเช่นในตัวอย่างที่ 3.7 ค่า λ เท่ากับ 0.674 ซึ่งใกล้เคียงกับ 0.5 จึงควรแปลงโดยการถอดรากแต่หากใช้โปรแกรม MINITAB แล้วโปรแกรมจะแปลงให้โดยอัตโนมัติ

3. หากค่า λ มีค่าใกล้เคียง 1 ไม่จำเป็นที่จะต้องแปลงข้อมูล

สรุป

การวิเคราะห์ความเหมาะสมของตัวแบบเป็นสิ่งที่จำเป็นในการวิเคราะห์การถดถอยเพื่อให้การวิเคราะห์ข้อมูลได้ถูกต้องและพยากรณ์ได้อย่างเที่ยงตรง ในการวิเคราะห์ความเหมาะสมสามารถทำได้หลายวิธี เช่น การใช้แผนภาพกระจายระหว่างตัวแปรหรือส่วนเหลือหรือค่าพยากรณ์ รวมไปถึงการวิเคราะห์โดยใช้วิธีทางสถิติที่มีความถูกต้องมากขึ้น หากข้อมูลมีรูปแบบที่ไม่เหมาะสมแล้วจำเป็นต้องมีการแปลงเพื่อให้ตัวแบบมีความถูกต้องมากขึ้น

คำถามท้ายบท

3.1 ข้อมูลราคาขายบ้าน (แสนบาท) กับยอดขายบ้านแบบต่างๆ (หลัง) ของบริษัทหนึ่งมีดังนี้

ราคา	134	67	44	39	29	20	17	19	7	10	5
ยอดขาย	42	100	61	60	79	99	91	127	115	250	260

(1) สร้างแผนภาพกระจายระหว่างตัวแปรทั้งสองพร้อมทั้งอธิบายความสัมพันธ์

(2) สร้างสมการถดถอยเชิงเส้นอย่างง่าย

(3) คำนวณค่าพยากรณ์ที่ราคาขายแต่ละค่า

(4) สร้างแผนภาพกระจายระหว่างส่วนเหลือกับค่าพยากรณ์พร้อมทั้งอธิบายแผนภาพที่ได้

(5) สร้างแผนภาพกระจายแบบโค้งปกติพร้อมทั้งอธิบายแผนภาพที่ได้

(6) คำนวณส่วนเหลือมาตรฐานของข้อมูลชุดนี้

3.2 จากข้อมูลในข้อ 2.13 จงสร้างแผนภาพดังต่อไปนี้

(1) แผนภาพกระจายระหว่างส่วนเหลือกับค่าพยากรณ์พร้อมทั้งอธิบายแผนภาพที่ได้

(2) แผนภาพกระจายแบบโค้งปกติพร้อมทั้งอธิบายแผนภาพที่ได้

3.3 จากข้อมูลข้างล่างจงตอบคำถามต่อไปนี้

X	50	62	34	63	81	97	36	25	40	70	74
Y	1.5	1.8	1.1	1.9	2.2	2.4	1.2	0.4	2.1	1.8	2.1

(1) สร้างสมการถดถอยเชิงเส้นอย่างง่าย

(2) สร้างแผนภาพกระจายแบบโค้งปกติพร้อมทั้งอธิบายแผนภาพที่ได้

(3) คำนวณส่วนเหลือปรับแล้วของข้อมูลชุดนี้

(4) สร้างแผนภาพกระจายระหว่างส่วนเหลือกับค่าพยากรณ์พร้อมทั้งอธิบายแผนภาพที่ได้

(5) หากท่านต้องการปรับปรุงตัวแบบท่านควรแปลงข้อมูลอย่างไร

3.4 จากข้อมูลในข้อ 3.3 จงทดสอบความเป็นอิสระของส่วนเหลือที่ได้โดยวิธี Durbin-Watson ที่ระดับนัยสำคัญ 0.05

3.5 จากข้อมูลในข้อ 3.3 จงใช้วิธี Modified Levene test ในการทดสอบความคงที่ของความแปรปรวนของส่วนเหลือที่ระดับนัยสำคัญ 0.05

3.6 ในการศึกษาความสัมพันธ์ระหว่างปริมาณเชื้อแบคทีเรียที่เหลือ (โคโลนี) กับเวลาที่ใช้ในการฆ่าเชื้อ (นาที) มีข้อมูลดังนี้

เวลา	178	110	94	83	70	50	50	30	29	18	15
จำนวน	1	2	3	4	5	6	7	8	9	10	11

(1) สร้างแผนภาพกระจายระหว่างตัวแปรทั้งสองพร้อมทั้งอธิบายความสัมพันธ์

(2) สร้างสมการถดถอยเชิงเส้นอย่างง่าย

(3) สร้างแผนภาพกระจายของส่วนเหลือกับเวลาพร้อมทั้งอธิบายแผนภาพที่ได้

(4) สร้างแผนภาพกระจายแบบโค้งปกติพร้อมทั้งอธิบายแผนภาพที่ได้

(5) หากท่านต้องการปรับปรุงตัวแบบท่านควรแปลงข้อมูลอย่างไร

3.7 จากข้อมูลข้างล่างจงตอบคำถามต่อไปนี้

X	1	0	0	2	1	3	0	1	2	0	2	1	3
Y	15	12	10	18	14	22	9	16	19	11	17	13	21

(1) สร้างแผนภาพกระจายระหว่างตัวแปรทั้งสองพร้อมทั้งอธิบายความสัมพันธ์

(2) สร้างสมการถดถอยเชิงเส้นอย่างง่าย

(3) สร้างแผนภาพกระจายของส่วนเหลือกับ X พร้อมทั้งอธิบายแผนภาพที่ได้

(4) สร้างแผนภาพกระจายของส่วนเหลือกับค่าพยากรณ์ ผลลัพธ์ที่ได้สอดคล้องกับข้อ (3) หรือไม่

(5) สร้างแผนภาพกระจายแบบโค้งปกติพร้อมทั้งอธิบายแผนภาพที่ได้

(6) ทดสอบความคงที่ของความแปรปรวนของส่วนเหลือโดยใช้วิธี Modified Levene test ที่ระดับนัยสำคัญ 0.05 แล้วเปรียบเทียบผลที่ได้กับแผนภาพในข้อ (3)

(7) ทดสอบความเหมาะสมของสมการถดถอยที่ระดับนัยสำคัญ 0.10

(8) หากท่านต้องการปรับปรุงตัวแบบท่านควรแปลงข้อมูลอย่างไร

3.8 ข้อมูลข้างล่างคือค่า X และส่วนเหลือที่ได้จากสมการถดถอยอย่างง่าย จงสร้างแผนภาพกระจายระหว่างค่าทั้งสองพร้อมทั้งอธิบายแผนภาพที่ได้ หากต้องการแปลงข้อมูลควรแปลงอย่างไร

X	5	6	7	8	9	10	11	12	13	14
ส่วนเหลือ	3.2	2.9	-1.7	-2.0	-2.3	-1.2	-0.9	0.8	0.7	0.5

3.9 จากข้อมูลข้างล่างจงสร้างแผนภาพกระจายระหว่างตัวแปรทั้งสองและทดสอบความเหมาะสมของสมการถดถอยที่ระดับนัยสำคัญ 0.05

X	6	6	6	8	8	10	10	14	14	22	22
Y	100	95	100	104	117	121	115	124	120	133	130

X	26	26	30	30	32	32
Y	180	135	149	146	150	153

3.10 หากท่านต้องการปรับปรุงสมการถดถอยของข้อ 3.9 ท่านจะทำอย่างไร