บทที่ 2

ตัวสถิติที่เกี่ยวข้องกับการวิจัย

การศึกษาวิธีตรวจสอบข้อมูลผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้น ในงานวิจัยครั้งนี้ เป็นการศึกษาในกรณีการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย (simple linear regression) เพื่อหาค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 (probability of type I error) และอำนาจการทดสอบ(power of the test) ของวิธีการทดสอบ 3 วิธีการคือ วิธีของทิตเจน, มัวร์ และเบคแมน (Tietjen, Moore and Beckman, 1973) วิธีของเมอวีน(Mervyn .G. Marasinghe, 1985) และวิธีของจีแบร์รี(G. Barrie Wetherill, 1986) โดยศึกษาค่าผิดปกติ 3 กรณีคือ กรณีเกิดข้อมูลผิดปกติ 1, 2 และ 3 ค่าตามลำดับ ซึ่งจะกล่าวถึงรายละเอียดของแต่ละวิธีการตรวจสอบข้อมูลผิดปกติต่อไป

รูปแบบของสมการถดถอยเชิงเส้นอย่างง่ายที่ศึกษาคือ

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \quad , \ i = 1, 2, 3, \ldots, n$$

ซึ่งเราสามารถเขียนสมการข้างต้นให้อยู่ในรูปของเมตริกซ์ได้ดังนี้

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

เมื่อ $\underline{y}$ คือ เวกเตอร์ของตัวแปรตามซึ่งมีขนาด nx1

X คือ เมตริกซ์คงที่ของตัวแปรอิสระขนาด nxp และมี rank = p

โดยที่ p คือ จำนวนพารามิเตอร์ที่ต้องการประมาณ

$\underline{\beta}$ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่ไม่ทราบค่าซึ่งมีขนาด px1

$\underline{\varepsilon}$ คือ เวกเตอร์ของความคลาดเคลื่อนซึ่งมีขนาด nx1

โดยมีข้อสมมติว่า $E(\underline{\varepsilon}) = \underline{0}$, $E(\underline{\varepsilon}\underline{\varepsilon}') = \sigma^2 I_n$

ซึ่งจากการประมาณค่าสัมประสิทธิ์การถดถอยโดยใช้วิธีกำลังสองน้อยที่สุด จะได้ว่า

(2.1) $$\underline{\beta} = (X'X)^{-1}X'\underline{y}$$

ดังนั้น

(2.2) $$\underline{\hat{y}} = X\underline{\beta}$$

แทนค่า $\underline{\beta}$ ของสมการ (2.1)ในสมการ (2.2) จะได้ว่า

(2.3) $\hat{y} = X(X'X)^{-1}X'y = Hy$

เมื่อ $H(nxn) = X(X'X)^{-1}X'$ เรียกว่า hat matrix เป็นเมตริกซ์ฉายา มีคุณสมบัติดังนี้

1. $H = H'$

และ 2. $H^2 = H$

ซึ่งจะมีบทบาทสำคัญในการศึกษาค่าผิดปกติและความคลาดเคลื่อน

(2.4) $\hat{\varepsilon} = y - \hat{y}$

แทนค่า $\hat{y}$ ของสมการ (2.3)ในสมการ (2.4)จะได้ว่า

$$\hat{\varepsilon} = y - Hy$$

$$\hat{\varepsilon} = (I-H)y$$

ซึ่งมีค่าประมาณความแปรปรวนของเวกเตอร์ความคลาดเคลื่อนคือ

$$\widehat{Var}(\hat{\varepsilon}) = (I-H)s^2$$

และค่าประมาณความแปรปรวนของความคลาดเคลื่อนที่ค่าสังเกตที่ i คือ

$$\widehat{Var}(\hat{\varepsilon}_i) = (1-h_{ii})s^2$$

เมื่อ h_{ii} เป็นสมาชิกแนวทแยงมุมของ hat matrix

และ s^2 เป็นตัวประมาณที่ไม่เอนเอียงของ σ^2 มีค่าเท่ากับ $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 / (n-p)$

จากการวิเคราะห์ความถดถอยเชิงเส้นข้างต้น สามารถนำมาหาค่าความคลาดเคลื่อนมาตรฐาน(standardized หรือ studentized residual) หรือความคลาดเคลื่อนที่ปรับแล้ว(adjusted residual) ได้ดังนี้

ความคลาดเคลื่อนมาตรฐาน(studentized residual , (R_i)) คือ

$$R_i = \hat{\varepsilon}_i / (s^2(1-h_{ii}))^{1/2}$$

และความคลาดเคลื่อนที่ปรับแล้ว(adjusted residual , (t_i)) คือ

$$t_i = \hat{\varepsilon}_i / (1-h_{ii})^{1/2}$$

ในการศึกษาค่าผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นค่า R_i จะเป็นตัวสำคัญในการบ่งชี้ค่าผิดปกติของวิธีการของทิตเจน, มัวร์ และเบคแมน และวิธีการของจีแบร์รี ส่วน t_i จะเป็นตัวสำคัญในการบ่งชี้ค่าผิดปกติของวิธีการของเมอวีน

รายละเอียดของแต่ละวิธีการมีดังต่อไปนี้

2.1 **วิธีการตรวจสอบของทิตเจน, มัวร์ และเบคแมน(TMB) (Tietjen, Moore and Beckman, 1973)**

เสนอโดย ทิตเจน, มัวร์ และเบคแมน เป็นการตรวจสอบค่าผิดปกติทีละค่าในข้อมูลแบบเรียงลำดับกัน(sequential data) โดยมีหลักเกณฑ์คือ ค่าสังเกตค่าใดที่ให้ค่าสูงสุดของค่าสัมบูรณ์ของค่าความคลาดเคลื่อนมาตรฐาน(maximum absolute studentized residual) ค่าสังเกตค่านั้นจัดว่าเป็นค่าผิดปกติ สำหรับตัวสถิติทดสอบของวิธีการนี้คือ ตัวสถิติทดสอบ R_n

$$R_n = \max_{i=1,2,..,n} |R_i|$$

เมื่อ R_i คือความคลาดเคลื่อนมาตรฐาน(studentized residual)

วิธีการตรวจสอบของทิตเจน, มัวร์ และเบคแมน มีขั้นตอนการคำนวณดังนี้

ก) คำนวณหาค่า R_i ทุกค่า $i = 1, 2, \ldots, n$

ข) คำนวณตัวสถิติทดสอบ R_n

ค) ตรวจสอบค่าสังเกตที่ตรงกับค่าสถิติทดสอบ R_n ว่าเป็นค่าผิดปกติหรือไม่ โดยมีสมมติฐานดังนี้

H_0 : ไม่มีค่าข้อมูลผิดปกติ

H_a : มีค่าข้อมูลผิดปกติหนึ่งค่า

โดยนำค่า R_n ที่คำนวณได้ เปรียบเทียบกับค่าขอบเขตวิกฤตจากตารางของลูนด์ (Lund, 1975)

ง) เกณฑ์การตัดสินใจกำหนดดังนี้ ถ้า R_n(คำนวณ) > R_n(ตาราง) จะปฏิเสธ H_0 แสดงว่าค่าสังเกตที่ i ซึ่งมีค่าตรงกับค่าสถิติทดสอบ R_n เป็นค่าผิดปกติ

จ) ในกรณีที่ปฏิเสธ H_0 จะตัดค่าสังเกตค่าที่ i ซึ่งตรงกับค่าสถิติทดสอบ R_n ออกไปแล้ววิเคราะห์ข้อมูลขนาด n-1 เช่นเดียวกับข้อ ก) ถึง ง) จะกระทำซ้ำจนกว่าจะยอมรับสมมติฐานว่าง H_0 จึงจะหยุดทำการทดสอบ

2.2 วิธีการตรวจสอบของเมอวิน จี มาราซิงห์(M) (Mervyn .G. Marasinghe, 1985)

วิธีการนี้เสนอโดยเมอวิน จี มาราซิงห์ ในปีค.ศ.1985 ซึ่งเป็นวิธีตรวจสอบค่าผิดปกติกรณีที่มีค่าผิดปกติหลายค่าในการวิเคราะห์ความถดถอยเชิงเส้น(Multistage procedure for detecting several outliers in linear regression) โดยมีหลักเกณฑ์ดังนี้ จะใช้ค่าความคลาดเคลื่อนที่ปรับแล้ว(adjusted residual, (ti)) เป็นตัวบ่งชี้ค่าผิดปกติ โดยนำค่าความคลาดเคลื่อนที่ปรับแล้วที่มีค่าสูงสุดในแต่ละรอบ(ซึ่งจำนวนรอบสูงสุดที่เป็นไปได้คือ k^* รอบ) จำนวน k^* ค่า โดยเปรียบเสมือนว่าค่าสังเกต k^* ค่าในเซตนี้เป็นค่าผิดปกติ จากนั้นจึงทำการคำนวณตัวสถิติทดสอบ F_{k*}

$$F_{k*} = (S - Q_{k*})/S$$

เมื่อ $S = \frac{(n-p)s^2}{n}$

$$s^2 = \sum_{i=1} \hat{\varepsilon}_i^{\,2} /(n-p)$$

$$Q_{k*} = \sum_{i=1}^{k^*} (t_i)^2$$

โดยที่ t_i คือค่าความคลาดเคลื่อนที่ปรับแล้วที่มีค่าสูงสุดรอบที่ i(maximum absolute adjusted residual)

และ k^* คือ จำนวนรอบสูงสุดที่เป็นไปได้ซึ่งเป็นเลขจำนวนเต็มมีค่าตั้งแต่ 2 ถึง 5 ซึ่งจะต้องกำหนดขึ้นโดย $k^* = k + 1$ เมื่อ k คือ จำนวนค่าผิดปกติ

วิธีการตรวจสอบของเมอวิน จี มาราซิงห์ มีขั้นตอนการคำนวณดังนี้

ก) หาเซตของค่าสังเกตซึ่งมี k^* ค่า โดยพิจารณาจากค่าสัมบูรณ์ของค่าความคลาดเคลื่อนที่ปรับแล้วสูงสุดในแต่ละรอบ(t_i)

ข) คำนวณค่า residual sum of square ในแต่ละรอบ

ค) คำนวณค่าสถิติทดสอบ F_{k*}

ง) ตรวจสอบค่าผิดปกติโดยมีสมมติฐานดังนี้

H_0 : ไม่มีข้อมูลผิดปกติ

H_1 : มีข้อมูลผิดปกติอย่างมากที่สุด k^* ค่า

โดยการนำค่า F_{k^*} ที่คำนวณได้ เปรียบเทียบกับค่าขอบเขตวิกฤตในตารางของเมอวีน (Mervyn .G. Marasinghe, 1985)

จ) เกณฑ์การตัดสินใจกำหนดดังนี้ ถ้า F_{k^*}(คำนวน) < F_{k^*}(ตาราง) จะปฏิเสธสมมติฐานว่าง(null hypothesis: H_0) แสดงว่าค่าสังเกตที่ i ซึ่งมีค่าตรงกับค่า t_i เป็นค่าผิดปกติ

ฉ) ในกรณีที่ปฏิเสธ H_0 จะตัดค่าสังเกตค่าดังกล่าวในข้อ จ) ออกไป แล้ววิเคราะห์ข้อมูลขนาด n-1 เช่นเดียวกับข้อ ค) ถึง จ) จะกระทำซ้ำจนกว่าจะยอมรับสมมติฐานว่าง H_0 จึงจะหยุดทำการทดสอบ

ช) สรุปผลการทดสอบพร้อมสรุปค่าผิดปกติ

2.3 วิธีการตรวจสอบของจีแบร์รี(GB) (G. Barrie Wetherill, 1986)

เป็นวิธีการตรวจสอบค่าผิดปกติที่ปรับปรุงมาจาก วิธีของทิตเจน, มัวร์ และเบคแมน(TMB)โดยทำการตรวจสอบค่าผิดปกติทีละค่าและมีหลักเกณฑ์ดังนี้ จะใช้ค่าความคลาดเคลื่อนมาตรฐานเป็นตัวบ่งชี้ค่าผิดปกติ ถ้าค่าความคลาดเคลื่อนมาตรฐานมีค่าสูงมากแสดงว่าตัวแปรตาม y เป็นค่าผิดปกติ ตัวสถิติทดสอบของวิธีการของจีแบร์รี(GB)คือ

$$d_i = (R)^2/(n-p)$$

วิธีการตรวจสอบของจีแบร์รี มีขั้นตอนการคำนวณดังนี้

ก) คำนวณค่าความคลาดเคลื่อนมาตรฐาน (R_i)

ข) เลือกค่าความคลาดเคลื่อนมาตรฐานที่สูงสุด R โดยที่

$$R = \max_{i=1,2,..,n} |R_i|$$

(ซึ่งก็คือสถิติทดสอบ R_n ในวิธีการของทิตเจน, มัวร์ และเบคแมน(TMB))

ค) คำนวณสถิติทดสอบ d_i ซึ่งจีแบร์รี สามารถพิสูจน์ได้ว่า d_i เป็นตัวสถิติที่มีการแจกแจงแบบเบตา(beta distribution) ซึ่งมีพารามิเตอร์เป็น (1/2 , (n-p-1)/2)

ง) ตรวจสอบว่าค่าสังเกตที่ i ซึ่งตรงกับค่า d_i เป็นค่าผิดปกติโดยการทดสอบสมมติฐาน ดังนี้

H_0 : ไม่มีข้อมูลผิดปกติ

กล่าวคือ $E(\underset{\sim}{y}) = X\underset{\sim}{\beta}$

เทียบกับ H_1 : มีข้อมูลผิดปกติหนึ่งค่า

กล่าวคือ $E(\underset{\sim}{y}) = X\underset{\sim}{\beta} + \underset{\sim}{\alpha}$

เมื่อ $\underset{\sim}{\alpha}$ คือ เวกเตอร์ของค่าผิดปกติซึ่งมีขนาด nx1

โดยนำค่า d_i ที่คำนวณได้ในข้อ ค) เปรียบเทียบกับค่าขอบเขตวิกฤตจากตารางของลูนด์(Lund, 1975) ซึ่งมีความสัมพันธ์ ดังนี้

$$d_0 = (R_0)^2/(n-p)$$

เมื่อ R_0 เป็นค่าขอบเขตวิกฤตที่เปิดได้จากตารางของลูนด์

จ) เกณฑ์การตัดสินใจกำหนดดังนี้ ถ้า $d_i > d_0$ จะปฏิเสธสมมติฐานว่าง H_0 แสดงว่า ค่าสังเกตที่ i ซึ่งตรงกับค่า d_i เป็นค่าผิดปกติ

ฉ) กรณีที่ปฏิเสธ H_0 จะตัดค่าสังเกตที่ i ออกแล้ววิเคราะห์ข้อมูลขนาด n-1 เช่นเดียวกับข้อ ก) ถึง จ) จะกระทำซ้ำเช่นนี้จนกว่าจะยอมรับสมมติฐานว่าง H_0