การวิเคราะห์การถดถอยอย่างง่าย

ฉัตรศิริ ปิยะพิมลสิทธิ์

แนวคิดเบื้องต้นของเส้นการถดถอยอย่างง่าย

เมื่อพิจารณาความสัมพันธ์ระหว่างตัวแปร 2 ตัว (X และ Y) ผู้วิจัยมักจะคิดถึงการคำนวณหาความสัมพันธ์ระหว่างตัวแปรทั้งสอง เช่น สัมประสิทธิ์สหสัมพันธ์ (The Pearson Product-Moment Correlation Coefficient : r_{xy}) แต่มีอีกวิธีหนึ่งคือการค้นหาความสัมพันธ์ระหว่างตัวแปรทั้งสองโดยผ่านการวิเคราะห์การถดถอยในรูปของการทำนาย นั่นคือ ความสามารถของตัวแปรหนึ่งสามารถทำนายตัวแปรหนึ่ง โดยปกติเราใช้สัญลักษณ์ X นิยามว่าเป็นตัวแปรอิสระ (Independent variable) หรือตัวแปรทำนาย (Predictor variable) และ Y นิยามว่าเป็นตัวแปรตาม (Dependent variable) หรือตัวแปรเกณฑ์ (Criterion variable)

ตัวอย่าง สมมติว่า ผู้บริหารมหาวิทยาลัยแห่งหนึ่งต้องการใช้คะแนนสอบเข้าระดับบัณฑิตศึกษา (Graduate Record Exam : GRE) ในการทำนายระดับผลการเรียนเฉลี่ย (Grade Point Average : GPA) เพื่อจะทำการตัดสินใจในการสรรหาเครื่องมือเพื่อสอบเข้าศึกษาต่อ ซึ่งผู้ที่ทำการศึกษาเรื่องนี้มีคำถามว่า GRE (Independent or Predictor variable) สามารถทำนายผลการเรียนเฉลี่ย (Dependent or Criterion variable) ได้มากน้อยเพียงใด? จากตัวอย่างนี้จะใช้การวิเคราะห์การถดถอยอย่างง่าย เพราะมีตัวแปรทำนายเพียงตัวเดียว

เส้นสมการถดถอยอย่างง่ายสามารถเขียนเป็นสมการเชิงเส้นตรงได้ดังนี้

$$Y = bX + a$$

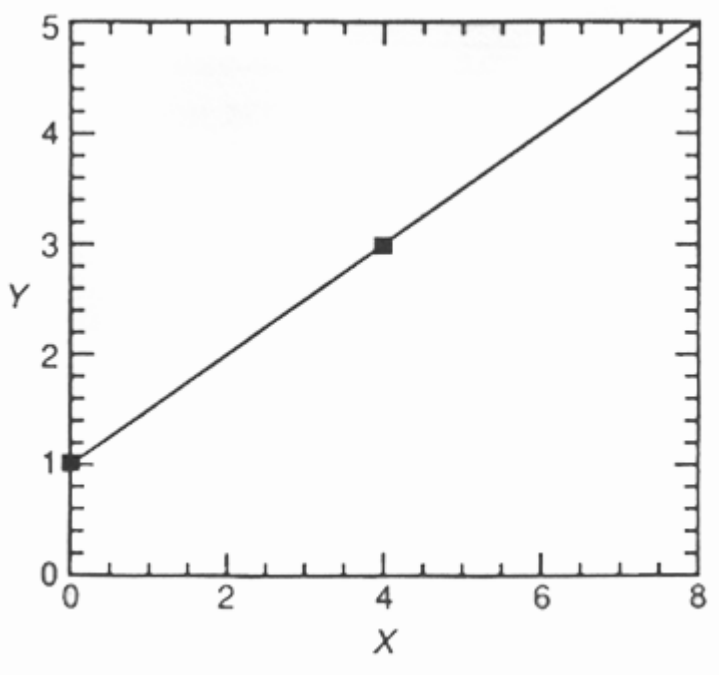
เมื่อ X (ตัวแปรทำนาย) ใช้ทำนายตัวแปร Y (ตัวแปรเกณฑ์) ความชันของเส้นจะใช้สัญลักษณ์ b และเป็นตัวบ่งชี้จำนวนหน่วยของ Y ที่เปลี่ยนแปลงไป การเปลี่ยนแปลงของ X หนึ่งหน่วยจะทำให้ Y เปลี่ยนแปลงไป b หน่วย จุดตัดของ Y จะใช้สัญลักษณ์ a ซึ่งเป็นจุดที่เส้นถดถอยตัดกับแกน Y และ a มีค่าเท่ากับ Y เมื่อ X มีค่าเป็นศูนย์

พิจารณาการสร้างเส้นตรง Y = 0.5X + 1.0 แสดงในภาพประกอบ 1 เราจะเห็นชัดว่าจุดตัดของเส้นอยู่ที่ Y = 1.0 ดังนั้นจุดตัดจึงเท่ากับ 1.0 ความชันของเส้นนิยามว่าเป็นการเปลี่ยนแปลงใน Y ที่หารด้วยการเปลี่ยนแปลงใน X

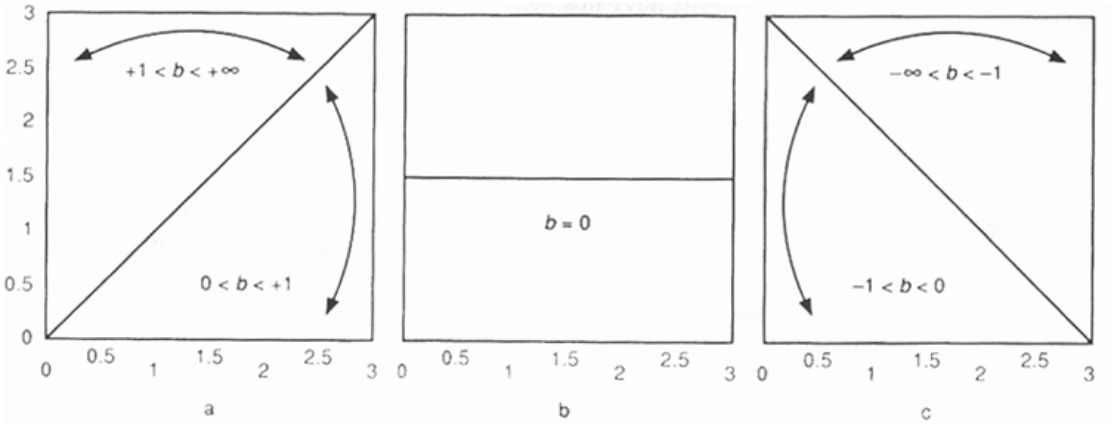
$$B = \Delta Y/\Delta X \quad \text{หรือ} \quad (Y_2 - Y_1)/(X_2 - X_1)$$

จากตัวอย่างนี้ จุดสองจุดที่แสดงในภาพประกอบ 1 (X_1, Y_1) และ (X_2, Y_2) ตกบนเส้นตรงที่ตำแหน่ง (0, 1) และ (4, 3) ตามลำดับ เราคำนวณความชันด้วย 2 จุดนี้จะได้ (3 – 1)/(4 - 0) = 0.5 ถ้าเราเลือกจุดสองจุดที่ตกบนเส้นตรง ณ ตำแหน่งอื่น ๆ ความชันที่คำนวณได้จะ

เท่ากับ 0.5 นั่นคือ 2 จุดใด ๆ บนเส้นตรงที่เลือก ค่าความชันจะได้เท่าเดิมเสมอ ค่าคงที่คือ 0.5 และสองจุดนี้เมื่อลากเส้นตรงผ่านแล้วจะเกิดเส้นตรงที่มีความชัน 0.5 และมีจุดตัดที่ 1.0 พิจารณาตัวอย่างของการพล็อตเส้นตรงในภาพประกอบ 2 ในภาพประกอบ 2(a) เส้นตรงแนวทแยงจะมีความชัน +1.00 ซึ่งเส้นนี้จะเป็นเส้นอ้างอิง เส้นที่ตกอยู่ในตำแหน่งทางซ้ายและอยู่เหนือเส้นของภาพประกอบ 2(a) จะบ่งชี้ว่ามีความชันอยู่ระหว่าง +1.00 และ +∞ เส้นที่ตกทางขวาและอยู่ใต้เส้นจะบ่งชี้ว่ามีความชันอยู่ระหว่าง 0 และ +1.00 ภาพประกอบ 2(a) บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร X และ Y เป็นบวก ในภาพประกอบ 2(b) ความชันจะเท่ากับ 0 ซึ่งเป็นเส้นที่ขนานกับแกน X และตั้งฉากกับแกน Y บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร X และ Y เป็นศูนย์ ในภาพประกอบ 2(c) ความชันจะมีค่า −1.00 และเส้นที่ตกอยู่ในตำแหน่งทางซ้ายและอยู่ใต้เส้นจะมีค่าระหว่าง −1.00 และ 0 เส้นที่ตกอยู่ในตำแหน่งทางขวาและอยู่เหนือเส้นจะมีค่าระหว่าง −1.00 และ −∞ ภาพประกอบ 2(c) บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร X และ Y เป็นลบ สังเกตว่าสัญลักษณ์ของความชัน (เป็นบวกหรือลบ) จะเหมือนกับสัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์ นั่นคือถ้าคู่ของ X และ Y เพิ่มขึ้น ความชันและสัมประสิทธิ์สหสัมพันธ์จะมีค่าเป็นบวก



ภาพประกอบ 1 เส้นสมการ Y = 0.5X + 1.0



ภาพประกอบ 2 ความชันของเส้นที่เป็นไปได้

สมการเส้นถดถอยอย่างง่ายสำหรับประชากร

จากแนวคิดของเส้นถดถอยในหัวข้อที่แล้ว เมื่อเรามีคะแนนแต่ละคนของกลุ่มประชากรทั้งตัวแปร X และ Y ซึ่ง X จะใช้ในการทำนาย Y ดังนั้น X จะนิยามว่าเป็นตัวแปรทำนาย และ Y เป็นตัวแปรเกณฑ์ ในทางกลับกันเมื่อใช้ Y ทำนาย X ผลที่ได้จะแตกต่างกันทั้งเส้นการถดถอย ความชันและจุดตัด ต่อไปเราจะนิยามสมการของเส้นถดถอยเท่ากับสมการเส้นตรง สมการถดถอยที่มี Y เป็นตัวแปรเกณฑ์ และ X เป็นตัวแปรทำนาย จะเรียกในภาษาทางสถิติว่า การถดถอย Y บน X

สมการถดถอยของประชากรสำหรับถดถอย Y บน X คือ

$$Y_i = \beta_{YX}X_i + \alpha_{YX} + \varepsilon_i$$

เมื่อ Y คือตัวแปรเกณฑ์, X คือตัวแปรทำนาย, β_{YX} คือความชันของเส้นถดถอยสำหรับ Y ที่ถูกทำนายด้วย X, α_{YX} จุดตัดของเส้นถดถอยสำหรับ Y ที่ถูกทำนายด้วย X, ε_i คือความคลาดเคลื่อนในการทำนาย (ส่วนของ Y_i ที่ไม่สามารถทำนายได้ด้วย X_i) และ i คือตัวบ่งชี้กลุ่มตัวอย่างแต่ละคน ตัวบ่งชี้ i สามารถมีค่าจาก 1 ถึง N เมื่อ N คือจำนวนของประชากร

ในทางกลับกัน ถ้า Y ใช้ในการทำนาย X แล้วความชันจะเขียนเป็น β_{XY} และ จุดตัดจะเขียนเป็น α_{XY} ดังนั้น ลำดับของตัวห้อยบ่งชี้ถึงตัวแปรเกณฑ์ (ตัวห้อยตัวแรก) และตัวแปรทำนาย (ตัวห้อยตัวที่สอง)

สมการทำนายสำหรับประชากรคือ

$$Y_i' = \beta_{YX}X_i + \alpha_{YX}$$

เมื่อ Y_i' คือค่าในการทำนาย Y เมื่อแทนค่า X ดังนั้นเราจะเห็นว่าค่าความคลาดเคลื่อนในการทำนายของประชากรนิยามได้ว่า

$$\varepsilon_i = Y_i - Y_i'$$

ความแตกต่างประการเดียวระหว่างสมการถดถอยกับสมการทำนายก็คือ สมการถดถอยจะรวมความคลาดเคลื่อนในการทำนายเข้าไว้ด้วย แต่ในสมการทำนายจะรวมความคลาดเคลื่อนในการทำนายอยู่ในส่วนของ Y'

พิจารณาสำหรับการประยุกต์ใช้ความแตกต่างระหว่างสมการการถดถอยและสมการทำนายที่รู้ค่า X และ Y และจะใช้สมการทำนายสำหรับกลุ่มประชากรภายหลังจากที่ใช้ทำนาย Y จากค่า X โดยจะใช้ตัวอย่าง GRE ในการพัฒนาสมการถดถอยสำหรับประชากรนักเรียนในมหาวิทยาลัยที่ศึกษาอยู่ในปัจจุบันที่มีการวัด GPA ซึ่งจะได้ความชันและจุดตัด ท้ายสุดก็จะได้สมการทำนายที่ใช้ในการตัดสินใจคัดเลือกนักเรียนในปีถัดไปโดยอาศัยคะแนน GRE เป็นฐาน

ความชันและจุดตัดของกลุ่มประชากรในเส้นถดถอยอย่างง่ายสามารถเขียนเป็นผลคูณดังนี้

$$\beta_{YX} = [N\Sigma XY - (\Sigma X)\Sigma Y]/[N\Sigma X^2 - (\Sigma X)^2]$$

และ

$$\alpha_{YX} = \mu_Y - \beta_{YX}\mu_X \ ,$$

เมื่อผลรวมจะเป็นการรวมค่าตั้งแต่ i = 1,...,N, μ_Y คือค่าเฉลี่ยของประชากรสำหรับ Y ($\mu_Y = \Sigma Y/N$), และ μ_X คือค่าเฉลี่ยของประชากรสำหรับ X ($\mu_X = \Sigma X/N$) , สังเกตว่าก่อนหน้านี้ใช้วิธีสำหรับคำนวณความชันและจุดตัดสำหรับเส้นตรงไม่ได้ใช้การวิเคราะห์การถดถอย ซึ่งเราจะอธิบายในหัวข้อต่อ ๆ ไป

สมการของเส้นถดถอยอย่างง่ายสำหรับกลุ่มตัวอย่าง

สมการการถดถอยที่ไม่เป็นมาตรฐาน

ถ้าเรากลับไปดูโลกแห่งความเป็นจริงของสถิติสำหรับกลุ่มตัวอย่าง ให้เราพิจารณาสมการเส้นถดถอยอย่างง่ายของกลุ่มตัวอย่าง โดยปกติ อักษรกรีกจะอ้างอิงถึงพารามิเตอร์ของประชากรและอักษรโรมันจะอ้างอิงถึงสถิติของกลุ่มตัวอย่าง สมการการถดถอยอย่างง่ายสำหรับการถดถอย Y บน X คือ

$$Y_i = b_{YX}X_i + a_{YX} + e_i \ ,$$

เมื่อ Y และ X คือตัวแปรเกณฑ์และตัวแปรทำนาย ตามลำดับ , b_{YX} คือความชันของเส้นถดถอยสำหรับ Y ที่ถูกทำนายด้วย X , a_{YX} คือจุดตัดของเส้นถดถอยสำหรับ Y ที่ถูกทำนายด้วย X , e_i คือความคลาดเคลื่อนในการทำนาย (ส่วนของ Y_i ที่ไม่สามารถทำนายได้ด้วย X_i) ดัชนี i สามารถมีค่าตั้งแต่ 1 ถึง n เมื่อ n คือขนาดของกลุ่มตัวอย่าง และเขียนได้ว่า i = 1,...,n

สมการทำนายของกลุ่มตัวอย่างคือ

$$Y_i' = b_{YX}X_i + a_{YX}$$

เมื่อ Y_i' คือค่าที่ถูกทำนายของ Y เมื่อรู้ค่า X ดังนั้นเราจะเห็นว่าความคลาดเคลื่อนในการทำนายของกลุ่มตัวอย่างจะเท่ากับ

$$e_i = Y_i - Y_i'$$

ความชันและจุดตัดสามารถคำนวณได้ว่า

$$b_{YX} = [N\Sigma XY - (\Sigma X)\Sigma Y]/[N\Sigma X^2 - (\Sigma X)^2]$$

และ

$$a_{YX} = \bar{Y} - b_{YX}\,\bar{X}$$

เมื่อผลรวมควรจะเป็นการรวมทั้ง i = 1,...,n , $\bar{Y}$ คือค่าเฉลี่ยของกลุ่มตัวอย่างของตัวแปร Y ($\bar{Y} = \Sigma Y/n$) , และ $\bar{X}$ คือค่าเฉลี่ยของกลุ่มตัวอย่างของตัวแปร X ($\bar{X} = \Sigma X/n$) ค่าความชันจะอ้างอิงในทางใดทางหนึ่งต่อไปนี้ 1) ค่าคาดหวังหรือค่าการทำนายของตัวแปร Y เมื่อ X เปลี่ยนแปลงไป 1 หน่วย 2) อิทธิพลของ X ต่อ Y และ 3) สัมประสิทธิ์การถดถอยของคะแนนดิบ ค่าจุดตัดจะอ้างอิงในทางใดทางหนึ่งต่อไปนี้ 1) จุดที่เส้นถดถอยตัดกับแกน Y 2) ค่าของ Y เมื่อ X เป็น 0 และ 3) ค่าเฉลี่ยของ Y เมื่อ X เป็น 0

พิจารณาตัวอย่างจริงที่จะนำมาใช้ในการวิเคราะห์ โดยใช้คะแนน GRE เป็นคะแนนที่ใช้ทำนายคะแนนสอบกลางภาค ซึ่ง GRE จะมีคะแนนอยู่ในช่วง 20 ถึง 80 และคะแนนสอบกลางภาคมีช่วงคะแนนอยู่ระหว่าง 0 ถึง 50 กลุ่มตัวอย่างเป็นนักเรียน 10 คน แสดงข้อมูลในตาราง 1 ต่อไปนี้จะแสดงการวิเคราะห์เส้นถดถอยอย่างง่าย โดยมีกลุ่มตัวอย่างที่สังเกตได้ (i = 1,...,10) ตัวแปรคะแนน GRE และตัวแปรคะแนนสอบกลางภาค ดังแสดงใน 3 สดมภ์แรกของตาราง ตามลำดับ

คะแนนเฉลี่ยของกลุ่มตัวอย่าง ซึ่งจะคำนวณคะแนนเฉลี่ยของ GRE ได้ $\bar{X} = 55.5$ และคะแนนสอบกลางภาค $\bar{Y} = 38$ ความชันและจุดตัดคำนวณได้ดังนี้

$$b_{YX} = [N\Sigma XY - (\Sigma X)\Sigma Y]/[N\Sigma X^2 - (\Sigma X)^2]$$
$$= [10(21905) - (555)(380)]/[10(32355) - (555)^2]$$
$$= 8150/15525$$
$$= 0.52$$

และ

$$a_{YX} = \bar{Y} - b_{YX}\,\bar{X}$$
$$= 38 - 0.52(55.5)$$
$$= 8.86$$

ตาราง 1 ข้อมูลสถิติของการสอบกลางภาค

นักเรียน	GRE	Midterm	Residual	Predicted Y
1	37	32	3.71	28.29
2	45	36	3.51	32.49
3	43	27	-4.44	31.44
4	50	34	-1.11	35.11
5	65	45	2.01	42.99
6	72	49	2.34	46.66
7	61	42	1.11	40.89
8	57	38	-0.79	38.79
9	48	30	-4.06	34.06
10	77	47	-2.29	49.29
Sums : $\Sigma X = 555$; $\Sigma Y = 380$; $\Sigma X^2 = 32355$; $\Sigma Y^2 = 14948$; $\Sigma XY = 21905$				

แปลความหมายค่าจุดตัดและความชันได้ดังนี้ ความชัน 0.52 หมายความว่า ถ้าคะแนน GRE เพิ่มขึ้น 1 หน่วยแล้วคะแนนสอบกลางภาคจะเพิ่มขึ้น 0.52 หน่วย คะแนนจุดตัด 8.86 หมายความว่าถ้าคะแนน GRE เป็น 0 แล้วคะแนนสอบกลางภาคจะเป็น 8.86 อย่างไรก็ตามมันเป็นไปไม่ได้ที่คะแนน GRE จะเป็น 0 เพราะว่าคะแนน GRE ต่ำสุดที่ควรจะได้ก็คือ 20

นำค่าที่ได้ทั้งหมดมาแทนค่าในสมการเส้นถดถอยอย่างง่ายของกลุ่มตัวอย่างได้

$$Y_i = b_{YX}X_i + a_{YX} + e_i ,$$
$$= 0.52X_i + 8.86 + e_i$$

ถ้าคะแนน GRE ได้ 63 คะแนนแล้วผลการทำนายคะแนนสอบกลางภาคจะได้

$$Y_i = 0.52(63) + 8.86$$
$$= 41.94$$

ดังนั้นสมการทำนายที่ได้นี้ควรจะทำนายได้ว่าคะแนนสอบกลางภาคจะได้ประมาณ 42 คะแนน

สมการการถดถอยที่เป็นมาตรฐาน

ทั้งหมดที่กล่าวมาข้างต้นเป็นการสร้างสมการถดถอยที่เกี่ยวข้องกับการใช้คะแนนดิบ ความชันที่ถูกประมาณค่าก็ไม่เป็นมาตรฐานหรือความชันที่อยู่ในรูปของคะแนนดิบเพราะว่ามันจะทำนายการเปลี่ยนแปลงของ Y ซึ่งเป็นหน่วยในรูปของคะแนนดิบ เมื่อ X ที่อยู่ในรูปของคะแนนดิบเปลี่ยนแปลงไป 1 หน่วย ในอีกกรณีหนึ่งเราอาจจะต้องการแสดงการถดถอยของ Y บน X ใน

หน่วยของคะแนนมาตรฐาน z มากกว่าหน่วยของคะแนนดิบ เมื่อ $z(X_i) = (X_i - \bar{X})/s_X$, และ $z(Y_i) = (Y_i - \bar{Y})/s_Y$ และ s_X และ s_Y คือส่วนเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่างสำหรับ X และ Y ตามลำดับ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวแปรที่เป็นมาตรฐานทั้งคู่ (z_X และ z_Y) ก็คือ 0 และ 1 ตามลำดับ สมการทำนายที่เป็นมาตรฐานของกลุ่มตัวอย่างจะกลายเป็น

$$z(Y_i') = b^*_{YX} z(X_i)$$
$$= r_{XY} z(X_i)$$

ความชันของการถดถอยมาตรฐาน b^*_{YX} เท่ากับ r_{XY} คือสหสัมพันธ์ระหว่าง X และ Y เมื่อ $z(Y_i')$ และ $z(X_i)$ เป็นคะแนนมาตรฐาน z ของตัวแปรเกณฑ์และตัวแปรทำนาย ตามลำดับ สัมประสิทธิ์สหสัมพันธ์ของกลุ่มตัวอย่างคำนวณได้ด้วยสูตร

$$r_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

ไม่มีค่าจุดตัดในสมการทำนายเพราะค่าเฉลี่ยของคะแนนมาตรฐาน z ของทั้ง X และ Y คือ 0 ($a_{YX} = \bar{Y} - b_{YX}\bar{X}$ ในรูปของคะแนนดิบ และ $a^*_{YX} = \bar{z}_Y - b^*_{YX}\bar{z}_X = 0$ ในรูปของคะแนนมาตรฐาน) โดยสรุป ความชันในรูปของคะแนนมาตรฐานเท่ากับสัมประสิทธิ์สหสัมพันธ์ และจุดตัดในรูปคะแนนมาตรฐานจะเท่ากับ 0

สำหรับตัวอย่างในตาราง 1 คำนวณสัมประสิทธิ์สหสัมพันธ์ได้ดังนี้

$$r_{XY} = \frac{10(21905) - (555)(380)}{\sqrt{[10(32355) - (555)^2][10(14948) - (380)^2]}}$$
$$= 0.92$$

ดังนั้น สมการทำนายในรูปของคะแนนมาตรฐานของกลุ่มตัวอย่างคือ

$$z(Y_i') = 0.92z(X_i)$$

ความชัน 0.92 แปลความหมายได้ว่า คะแนนสอบกลางภาคควรจะเพิ่มขึ้น 0.92 หน่วย เมื่อคะแนน GRE เพิ่มขึ้น 1 หน่วย จำนวน 1 หน่วยที่เพิ่มขึ้นจะเหมือนกับการเพิ่มขึ้น 1 ส่วนเบี่ยงเบนมาตรฐาน เพราะว่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงคะแนน z จะมีค่าเท่ากับ 1

ในสถานการณ์ใดที่ต้องใช้การวิเคราะห์การถดถอยที่เป็นมาตรฐานหรือไม่เป็นมาตรฐาน นักสถิติอย่าง Pedhazur (1982) อธิบายว่า b^* สามารถยืดหยุ่นได้เมื่ออ้างอิงไปยังกลุ่มตัวอย่างอื่น ตัวอย่างเช่น อ้างอิงไปยังมหาวิทยาลัยแห่งอื่น ๆ b^* ควรจะใช้ได้กับกลุ่มตัวอย่างที่แตกต่างกัน นักวิจัยส่วนมากจึงนิยมใช้ b มากกว่าในการเปรียบเทียบผลการทำนายของตัวแปรในกลุ่มตัวอย่างหรือประชากรที่แตกต่างกัน

ความคลาดเคลื่อนในการทำนาย

ก่อนหน้านี้ได้อ้างอิงถึงการทำนาย Y จาก X แต่การทำนายได้อย่างสมบูรณ์จะเกิดขึ้นเมื่อ ความสัมพันธ์ระหว่างตัวแปรทำนายและตัวแปรเกณฑ์เป็นไปอย่างสมบูรณ์ ($r_{XY} = \pm 1.0$) เมื่อพัฒนาสมการถดถอย เรารู้ค่าของ Y ความชันและจุดตัดที่เกิดจากการประมาณค่า เราจะใช้สมการทำนายนี้ทำนาย Y จาก X เมื่อเราไม่รู้ค่า Y เราจะนิยามค่าที่ถูกทำนายของ Y ด้วย Y' ค่าที่ถูกทำนาย Y' สามารถคำนวณได้เมื่อรู้ค่า X ในสมการทำนาย ถ้า $Y'_i = Y_i$ สำหรับคนที่ i แล้วจะเป็นการทำนายอย่างสมบูรณ์ (Perfect prediction) อย่างไรก็ตาม

เราสามารถคำนวณค่า Y' ของแต่ละคนจากสมการทำนายและเปรียบเทียบค่า Y ที่ได้จริง เราจะได้ค่าความคลาดเคลื่อนในการทำนาย

$$e_i = Y_i - Y'_i$$

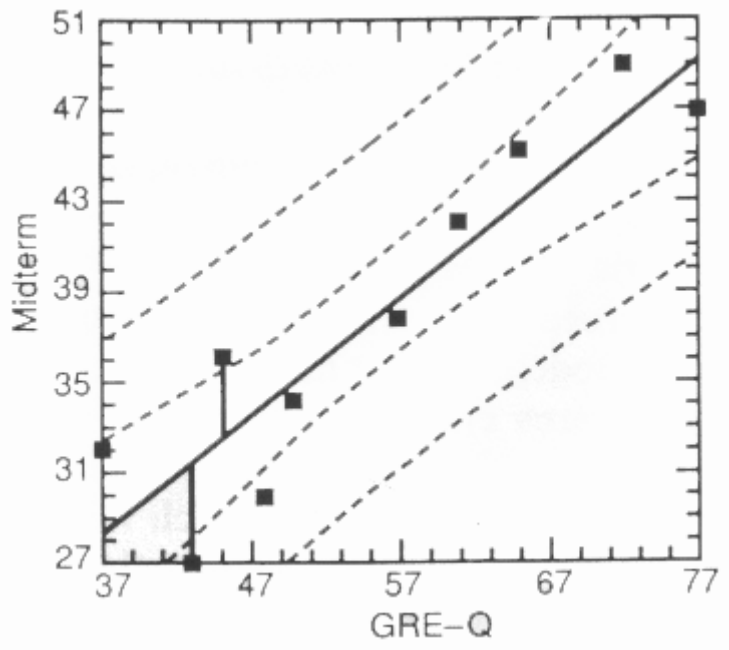
สำหรับ i = 1,...,n โดยที่ e_i คือค่าความคลาดเคลื่อนที่ถูกประมาณค่า หรือความคลาดเคลื่อนในการทำนาย และส่วนของ Y_i ที่ไม่สามารถคำนวณได้จาก X_i เทอมของความคลาดเคลื่อนนี้จะมีค่าอย่างสุ่มในแต่ละบุคคล

ค่าความคลาดเคลื่อนและค่าที่ถูกทำนายเราจะใช้ในการคำนวณคะแนนสอบกลางภาค ซึ่งแสดงใน 2 สดมภ์สุดท้ายของตาราง 1 ตามลำดับ พิจารณาคนที่ 2 เมื่อคะแนนสังเกต GRE คือ 45 และคะแนนสอบกลางภาคจริงได้ 36 คะแนนทำนายสอบกลางภาคได้ 32.49 และความคลาดเคลื่อนคือ +3.51 บ่งชี้ว่าคนที่ 2 มีคะแนนสอบกลางภาคที่สังเกตได้สูงกว่าคะแนนสอบกลางภาคที่ทำนายด้วย GRE ค่าความคลาดเคลื่อนในการทำนาย จะมีค่าเป็นบวก ถ้าความคลาดเคลื่อนในการทำนายเป็นลบ (เช่นคนที่ 3) บ่งชี้ว่า คะแนนที่สังเกตได้น้อยกว่าคะแนนที่ถูกทำนาย พิจารณาคนที่ 3 GRE ได้ 43 คะแนนสอบกลางภาคได้ 27 คะแนนที่ถูกทำนายคือ 31.44 ดังนั้นความคลาดเคลื่อนในการทำนายคือ −4.44 จะเห็นว่า คนที่ 2 คะแนนสังเกตได้สูงกว่าคะแนนทำนาย คนที่ 3 คะแนนที่สังเกตได้ต่ำกว่าคะแนนทำนาย อาจเป็นไปได้ว่าบุคคลทั้งสองมีผลสัมฤทธิ์สูงเกินจริง (overachiever) และผลสัมฤทธิ์ต่ำเกินจริง (underachiever) ตามลำดับ

การถดถอยคะแนนสอบกลางภาคด้วย GRE แสดงได้ดังแผนภาพกระจัดกระจายในภาพประกอบ 3 ในภาพประกอบ 3 จะแสดงตำแหน่งของคะแนนแต่ละคนในคู่ของ X และ Y เส้นดำในแนวทแยงคือเส้นถดถอย สังเกตคนที่ 2 แสดงในตำแหน่งของ GRE = 45 และกลางภาค = 36 จะอยู่ห่างจากเส้นถดถอยเท่ากับขนาดของความคลาดเคลื่อนในการทำนาย โดยคนที่ 2 จะอยู่เหนือเส้นถดถอย และคะแนนกลางภาคสูงกว่าคะแนนทำนาย และมีความคลาดเคลื่อนเป็นบวกซึ่งตกอยู่เหนือเส้นถดถอย 3.51 หน่วย คนที่ 2 คะแนน GRE = 43 และกลางภาค 27 สังเกตว่าจะอยู่ตำแหน่งใต้เส้นถดถอย เพราะมีคะแนนกลางภาคต่ำกว่าคะแนนทำนาย และความคลาดเคลื่อนติดลบ ซึ่งตกอยู่ใต้เส้นถดถอย 4.44 หน่วย จุดที่ตกอยู่เหนือเส้นถดถอยมีความ

คลาดเคลื่อนในการทำนายเป็นบวก จุดที่ตกอยู่ใต้เส้นถดถอยมีความคลาดเคลื่อนในการทำนายเป็นลบ

ภาพประกอบ 3 การถดถอยของคะแนนสอบกลางภาคบน GRE



ถ้าเราพิจารณาสดมภ์ของความคลาดเคลื่อนในการทำนาย ในตาราง 1 จะเห็นว่าครึ่งหนึ่งเป็นบวก ครึ่งหนึ่งเป็นลบ แม้ว่าอาจจะไม่ได้เป็นแบบนี้เสมอไป ในทำนองเดียวกันภาพประกอบ 3 จำนวนจุดครึ่งหนึ่งอยู่เหนือเส้นถดถอย อีกครึ่งหนึ่งอยู่ใต้เส้นถดถอย ดังนั้นค่าเฉลี่ยของความคลาดเคลื่อนในการทำนายจะเป็น 0 เสมอ ($\bar{e} = 0$) ซึ่งผลนี้จะเป็นจริงเสมอที่ว่า ค่าเฉลี่ยของคะแนนที่สังเกตได้จะเท่ากับค่าเฉลี่ยของคะแนนทำนาย ($\bar{Y} = \bar{Y}' = 38$ สำหรับตัวอย่างนี้)

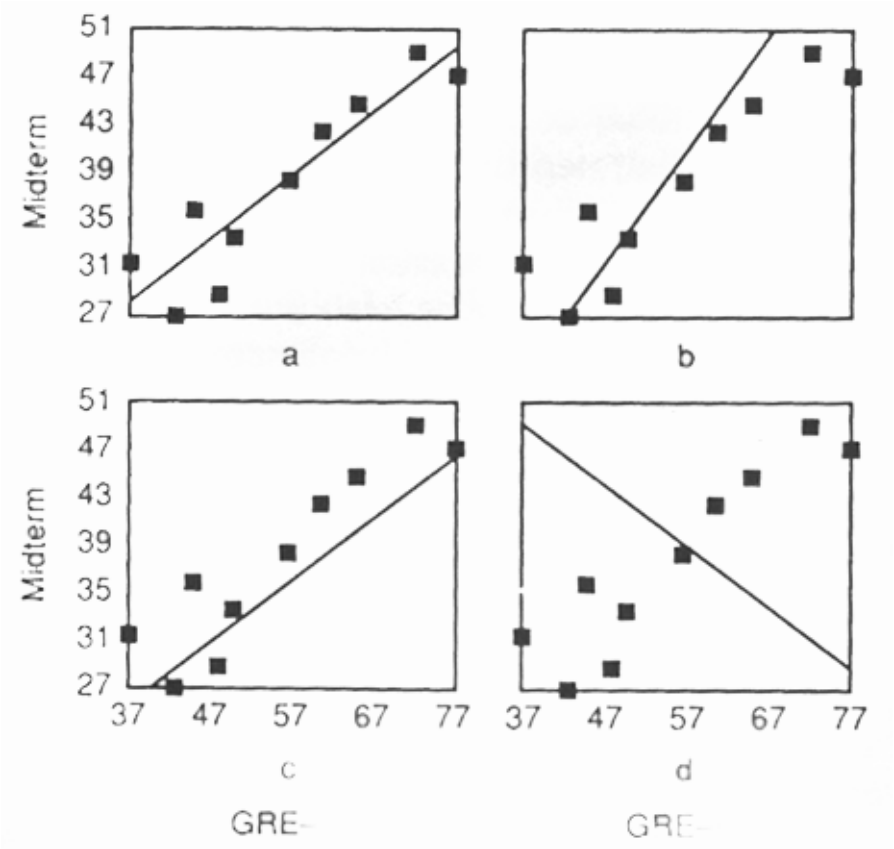
เกณฑ์กำลังสองต่ำสุด (Least Square Criterion)

มีวิธีการหนึ่งที่เลือกเข้ามาใช้ในการคำนวณหาความชันและจุดตัดของเส้นตรง หรือก็คือเกณฑ์ทางสถิติที่นำมาใช้ในการพิจารณาคำนวณหาค่า b และ a คือเกณฑ์กำลังสองต่ำสุด (least square criterion) ก็คือผลรวมกำลังสองของความคลาดเคลื่อนในการทำนายจะต้องต่ำสุด นั่นคือถ้าเราได้เส้นถดถอยจากความชันและจุดตัดแล้ว ค่าผลรวมของความคลาดเคลื่อนในการทำนายเมื่อยกกำลังสองแล้วจะต้องได้ค่าต่ำสุด

โดยสรุปแล้ว เกณฑ์กำลังสองต่ำสุด จะให้เราพิจารณาความชันและจุดตัดและเส้นถดถอยที่ผลรวมของความคลาดเคลื่อนในการทำนายยกกำลังสองแล้วได้ค่าต่ำสุด เรากล่าวไปแล้วถึงวิธีการคำนวณความชันและจุดตัดด้วยการประมาณค่ากำลังสองต่ำสุด และ b และ a เป็นการประมาณค่าจากกลุ่มตัวอย่างของพารามิเตอร์ β และ α ที่อ้างอิงด้วยวิธีกำลังสองต่ำสุด อาจจะได้ว่า ค่าเฉลี่ยเมื่อผลรวมของส่วนเบี่ยงเบนมาตรฐานของคะแนนดิบจากค่าเฉลี่ยน้อยกว่าผลรวมของส่วนเบี่ยงเบนมาตรฐานของคะแนนดิบจากค่าอื่น ๆ ดังนั้นวิธีการประมาณค่าเฉลี่ยก็คือการวัดแนวโน้มเข้าสู่ส่วนกลางที่เลือกมาใช้เป็นพื้นฐานของเกณฑ์กำลังสองต่ำสุด

มีตัวอย่างของเกณฑ์มากมายในข้อมูล ซึ่งจะแสดงด้วยภาพประกอบ 4 ซึ่งเป็นแผนภาพกระจัดกระจายของ GRE กับคะแนนกลางภาค แต่ละภาพจะมีเส้นถดถอยต่าง ๆ กัน (แตกต่างทั้งความชันและจุดตัด) ภาพประกอบ 4(a) เกิดจากการใช้เกณฑ์กำลังสองต่ำสุด ภาพประกอบ 4(b) เกิดจากการใช้เกณฑ์กำลังสองที่มากกว่า ส่วนภาพที่เหลือเกิดจากการใช้เกณฑ์อื่น ๆ ในการคำนวณผลรวมของความคลาดเคลื่อนในการทำนายยกกำลังสองในแต่ละเส้นสมการถดถอย ในตาราง 2 แสดงว่าเส้นถดถอยที่ใช้กำลังสองต่ำสุดมีค่าน้อยที่สุด ดังนั้นสำหรับข้อมูลชุดนี้และโดยทั่ว ๆ ไปเกณฑ์กำลังสองต่ำสุดเป็นสถิติที่ดีกว่าเกณฑ์อื่น ๆ

ภาพประกอบ 4 ตัวอย่างเส้นถดถอยที่แตกต่างกัน



ตาราง 2 ผลรวมกำลังสองของความคลาดเคลื่อนสำหรับเส้นถดถอยที่แตกต่างกัน

เส้น	ผลรวมกำลังสองของความคลาดเคลื่อน
a	80.16
b	833.00
c	281.75
d	1481.75

สัดส่วนของความแปรปรวนในการทำนาย (สัมประสิทธิ์การอธิบาย)

จากที่นำเสนอไปแล้ว เกิดคำถามว่า " ตัวแปรเกณฑ์ Y ถูกทำนายด้วยตัวแปรทำนาย X ได้ดีแค่ไหน" ในตัวอย่างนี้ จะสนใจว่า คะแนน GRE สามารถทำนายคะแนนสอบกลางภาคได้ดีแค่ไหน เราจะอธิบาย 2 สถานการณ์ที่เป็นไปได้ในตัวอย่างนี้คือ 1) ถ้า GRE พบว่าเป็นตัวแปรที่สามารถทำนายคะแนนสอบกลางภาคได้ดี ควรจะใช้ GRE มาเป็นเครื่องมือในการประเมินระดับทักษะของผู้เรียนแต่ละคน เช่น ผู้เรียนคนใดได้คะแนน GRE ต่ำก็อาจจะพัฒนาเอาใจใส่ให้เขาพัฒนาทักษะของตัวเองให้มากขึ้น 2) ถ้าพบว่า GRE ไม่สามารถทำนายสอบกลางภาคได้ดี อาจจะไม่ต้องใช้ GRE มาประเมินอีก และอาจค้นหาตัวแปรทำนายอื่น ๆ ที่สามารถทำนายคะแนนกลางภาคได้ดีกว่านี้

เราจะอธิบายประโยชน์ของตัวแปรทำนายได้อย่างไร วิธีง่ายที่สุดคือการแบ่งส่วนผลรวมกำลังสองของ Y ซึ่งใช้สัญลักษณ์ว่า SS_Y โดยใช้การวิเคราะห์ความแปรปรวน

เราสามารถแบ่งส่วนของ SS_Y ได้ว่า

$$SS_Y = SS_{reg} + SS_{res}$$

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y' - \bar{Y})^2 + \Sigma(Y - Y')^2$$

เมื่อ SS_Y คือผลรวมกำลังสองของ Y, SS_{reg} คือผลรวมกำลังสองของการถดถอย Y บน X (เขียนได้อีกรูปหนึ่งว่า $SS_{Y'}$), SS_{res} คือผลรวมกำลังสองของความคลาดเคลื่อน ผลรวมในที่นี้ก็คือ i = 1,..,n ในเทอมของ SS_Y จะแสดงความแปรปรวนรวมในคะแนนสังเกต Y ที่ไม่สามารถทำนายได้ด้วย X (ความแปรปรวนของความคลาดเคลื่อน) ในการคำนวณสามารถเขียนสูตร SS_Y, SS_{reg}, และ SS_{res} ได้ดังนี้

$$SS_Y = [n\Sigma Y^2 - (\Sigma Y)^2]/n$$

$$SS_{reg} = \frac{\{[n\Sigma XY - (\Sigma X)(\Sigma Y)]/n\}^2}{\{n\Sigma X^2 - (\Sigma X)^2]/n\}}$$

$$SS_{res} = SS_Y - SS_{reg}$$

ในขั้นสุดท้าย อัตราส่วนของความแปรปรวนในการทำนายต่อความแปรปรวนรวม สามารถเขียนในรูปทั่วไปได้ว่า

$$SS_{reg}/SS_Y = r^2_{XY}$$

อัตราส่วนนี้ เรียกว่าสัดส่วนของความแปรปรวนทั้งหมดใน Y ที่สามารถใช้สมการถดถอยได้ อัตราส่วนนี้เท่ากับ r^2_{XY} หรือกำลังสองของสัมประสิทธิ์สหสัมพันธ์เพียร์สัน และโดยทั่วไปจะอ้างอิงว่าเป็นสัมประสิทธิ์การอธิบาย

โดยทั่วไปไม่มีเกณฑ์ใด ๆ บอกขนาดของสัมประสิทธิ์การอธิบายว่าเท่าใดถึงจะบ่งชี้ว่าตัวแปรทำนายสามารถทำนายตัวแปรเกณฑ์ได้อย่างดี

จากข้อมูลตัวอย่าง การทำนายคะแนนสอบกลางภาคด้วยคะแนน GRE สามารถแสดงผลการคำนวณได้ดังนี้

$$SS_Y = [n\Sigma Y^2 - (\Sigma Y)^2]/n$$
$$= [10(14948) - (380)^2]/10$$
$$= 508.00$$

ถัดจากนั้นเราคำนวณหา SS_{reg} และ SS_{res} ได้ดังนี้

$$SS_{reg} = \frac{\{[n\Sigma XY - (\Sigma X)(\Sigma Y)]/n\}^2}{\{[n\Sigma X^2 - (\Sigma X)^2]/n\}}$$
$$= \frac{\{[10(21905) - (555)(380)]/10\}^2}{\{[10(32355) - (555)^2]/10\}}$$
$$= 427.84$$

และ $$SS_{res} = SS_Y - SS_{reg}$$
$$= 508.00 - 427.84$$
$$= 80.16$$

ต่อมาคำนวณใช้ตัวเศษเป็น SS_{reg} ซึ่งก็คือ SS_X เท่ากับ 1552.50 และขั้นตอนสุดท้ายคำนวณหาอัตราส่วน SS_{reg} ต่อ SS_Y หรือเขียนในรูปทั่วไปว่า

$$SS_{reg}/SS_Y = r^2_{XY}$$

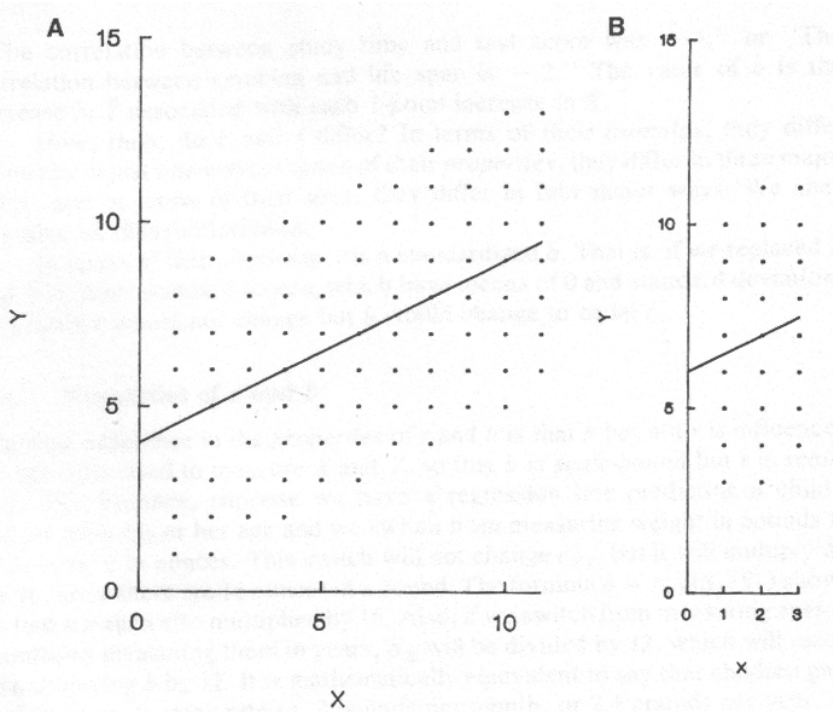
ผลสุดท้ายให้ตรวจสอบผลของข้อมูลตัวอย่างในการใช้การแบ่งส่วนของผลรวมกำลังสอง พบว่าสัมประสิทธิ์การอธิบายเท่ากับ 0.84 ถ้าเรานำค่าสหสัมพันธ์ 0.92 ยกกำลังสองจะได้ 0.84 ดังนั้น GRE ทำนายคะแนนสอบกลางภาคได้ประมาณ 84 เปอร์เซ็นต์ การทดสอบนัยสำคัญทางสถิติของค่าสัมประสิทธิ์การอธิบายจะกล่าวในหัวข้อถัดไป

คุณสมบัติของ r และ b

คุณสมบัติประการแรกที่แตกต่างกันของ r และ b ก็คือ b จะอ้างอิงอยู่กับหน่วยของการวัดในตัวแปร X และ Y ดังนั้น b จะเป็น scale-bound แต่ r จะเป็น scale-free เช่น สมมติว่าเรามีเส้นถดถอยที่ทำนายน้ำหนักของเด็กจากอายุ และวัดน้ำหนักเป็นปอนด์ แต่เปลี่ยนน้ำหนักจากเดิมวัดเป็นปอนด์มาเป็นออนซ์ ค่า r_{XY} จะไม่เปลี่ยน แต่ค่า b จะเปลี่ยนไปโดย 16 ออนซ์ = 1 ปอนด์ ดังนั้นค่า b จะเปลี่ยนไปโดยการคูณด้วย 16 ในอีกกรณีหนึ่งถ้าเราเปลี่ยนการวัดอายุของเด็ก จากเดิมวัดเป็นเดือน เปลี่ยนมาเป็นปี ซึ่ง 12 เดือน = 1 ปี ดังนั้นค่า b จะเปลี่ยนไปโดยการคูณด้วย 12 ดังนั้นอาจพูดได้ว่า เด็กมีน้ำหนักเพิ่มขึ้นโดยเฉลี่ย 0.2 ปอนด์ต่อเดือน หรือ 2.4 ปอนด์ต่อปี หรือ 38.4 ออนซ์ต่อปี หรือ 3.2 ออนซ์ต่อเดือน

คุณสมบัติประการที่สองที่แตกต่างกันของ r และ b ก็คือ ค่า r จะเพิ่มขึ้นเมื่อพิสัยของข้อมูลในตัวแปรเพิ่มขึ้น ขณะที่ b ยังคงที่ ตัวอย่างเช่น สมมติว่า X คือรายได้ Y คือเงินออม และมีความสัมพันธ์กันเป็นเส้นตรงระหว่าง X กับ Y ถ้าพิจารณาดี ๆ ในการศึกษา 2 ชิ้นคือชิ้น A และชิ้น B ที่อธิบายถึงความสัมพันธ์ระหว่าง X และ Y แต่การศึกษาชิ้น A จะเป็นการศึกษากับคนทั้งประเทศ ขณะที่ชิ้น B จะเป็นการศึกษาเฉพาะในเขตเมือง จึงเป็นไปได้ว่า การศึกษาความสัมพันธ์ระหว่างตัวแปร X และ Y ทั้งสองชิ้นนี้จะพบว่ามีค่า b เหมือนกัน แต่ชิ้น A จะมีค่า r สูงกว่าเพราะ X มีพิสัยกว้างกว่าการศึกษาในชิ้น B

ภาพประกอบ 5 พิสัยของตัวแปร X ที่ส่งผลต่อค่า r แต่ไม่ส่งผลต่อค่า b



ภาพประกอบ 5 แสดงการพล็อตจุด โดยในภาพประกอบ 5(b) แสดงความสอดคล้องของจุดกึ่งกลางของสดมภ์ทั้ง 3 สดมภ์ ในทั้งสองส่วนนี้ เส้นทแยงจะลากผ่านค่าเฉลี่ยทั้งหมด ดังนั้นเรารู้ได้โดยไม่ต้องคำนวณว่าคือเส้นถดถอย เส้นการถดถอยมีความชันเท่ากันในทั้งสองภาพ ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองในภาพประกอบ 5(a) มีค่า 5.94 และในภาพประกอบ 5(b) มีค่า 5.74 แต่ค่าสหสัมพันธ์ในภาพประกอบ 5(a) และ 5(b) จะมีค่า r_{XY} เป็น 0.54 และ 0.16 ตามลำดับ

ความแตกต่างประการที่สามในคุณสมบัติของ r และ b ก็คือ เมื่อ Y ได้รับอิทธิพลจากตัวแปรอื่น ๆ ที่ไม่สัมพันธ์กับ X ค่า r_{XY} จึงต่ำกว่าความเป็นจริง แต่จะไม่มีผลกับ b ค่า r_{XY} ในที่จริงแล้วเป็นการวัดที่มีความสำคัญของความสัมพันธ์ระหว่างตัวแปร XY ความสัมพันธ์ในองค์ประกอบอื่น ๆ ที่มีผลกับ X ขณะที่ b นั้นมีค่าการวัดเป็นขนาดสัมบูรณ์โดยเพิกเฉยต่อองค์ประกอบอื่น ๆ ยกตัวอย่างเช่น ถ้านานาชาติต้องการให้รณรงค์ในเรื่องความปลอดภัยโดยการลดอัตราการตายจากอุบัติเหตุ ซึ่งจะสันนิษฐานได้ว่า ไม่มีผลต่อความชันของเส้นถดถอย ที่อ้างอิงว่า ชีวิตจะสั้นลง 5 นาที ต่อบุหรี่ 1 มวน แต่การลดอัตราการตายจากอุบัติเหตุควรจะไปเพิ่มความสัมพันธ์ระหว่างการสูบบุหรี่กับความยืนยาวของชีวิต (นั่นคือทำให้ความสัมพันธ์เป็นลบมาก) เพราะมันจะไปเพิ่มความสำคัญของการสูบบุหรี่ที่สัมพันธ์กับองค์ประกอบอื่น ๆ อันจะมีผลต่อความยืนยาวของชีวิต

การทดสอบนัยสำคัญและช่วงความเชื่อมั่น

ในหัวข้อนี้จะอธิบาย 5 วิธีการทดสอบที่ใช้กับการถดถอยอย่างง่าย โดย 3 วิธีแรกเป็นการทดสอบนัยสำคัญทางสถิติที่เกี่ยวข้อง และอีก 2 วิธีสุดท้ายเป็นเทคนิคการหาช่วงความเชื่อมั่น

การทดสอบนัยสำคัญของ r^2_{XY} (Test of Significance of r^2_{XY})

การทดสอบแรกเป็นการทดสอบนัยสำคัญของ r^2_{XY} (หรือก็คือการทดสอบสัดส่วนของความแปรปรวนใน Y ที่ทำนายได้ด้วย X) สมมติฐานศูนย์และสมมติฐานอื่น ๆ สามารถเขียนได้ดังนี้

$$H_0 : \rho^2_{XY} = 0$$
$$H_1 : \rho^2_{XY} > 0$$

สามารถทดสอบนัยสำคัญได้ด้วยสถิติดังนี้

$$F = \frac{[r^2 / m]}{[(1 - r^2)/(n - m - 1)]}$$

เมื่อ F เป็นสถิติทดสอบ, r^2 คือสัมประสิทธิ์การอธิบาย (คือ r^2_{XY} หรือสัดส่วนของความแปรปรวนใน Y ที่ถูกทำนายได้ด้วย X), $1 - r^2$ คือสัดส่วนของความแปรปรวนใน Y ที่ไม่สามารถ

ทำนายได้ด้วย X, m คือจำนวนของตัวแปรทำนาย (ซึ่งในกรณีของการวิเคราะห์การถดถอยอย่างง่าย จะมีตัวแปรทำนายเพียง 1 ตัวเท่านั้น) และ n คือจำนวนกลุ่มตัวอย่าง สถิติ F-test ที่คำนวณได้จะนำไปเปรียบเทียบกับค่าวิกฤติ F ซึ่งจะเป็นการทดสอบแบบทางเดียว (one-tailed) ตามระดับนัยสำคัญที่กำหนด ที่มีองศาแห่งความเป็นอิสระ (degree of freedom) m และ n - m – 1 นำมาจากตาราง F นั่นคือ ค่าวิกฤติของ $_{(1-\alpha)}F_{m,(n-m-1)}$

สำหรับตัวอย่างคะแนนสอบกลางภาคและคะแนน GRE ของเรา สามารถทดสอบนัยสำคัญทางสถิติได้ดังนี้

$$F = \frac{[r^2/m]}{[(1-r^2)/(n-m-1)]}$$

$$= \frac{[0.84/1]}{[(1-0.84)/(10-1-1)]}$$

$$= 42.70$$

ค่าวิกฤติที่ระดับนัยสำคัญ .05 คือ $_{.95}F_{1,8} = 5.32$ ผลของการทดสอบทางสถิติปรากฏว่าค่า F ที่คำนวณได้มากกว่าค่าวิกฤติ F ที่เปิดจากตาราง นั่นคือปฏิเสธ H_0 ยอมรับ H_1 สรุปว่า ρ^2_{XY} มีค่าไม่เท่ากับศูนย์ ที่ระดับนัยสำคัญ .05 (นั่นคือคะแนน GRE สามารถทำนายสัดส่วนของความแปรปรวนของคะแนนสอบกลางภาคได้อย่างมีนัยสำคัญทางสถิติ)

การทดสอบนัยสำคัญของผลรวมกำลังสอง (Test of Significance of Sum of Squares)

การทดสอบอันดับที่ 2 เป็นการทดสอบสัดส่วนของผลรวมกำลังสองใน Y ที่ถูกทำนายด้วย X สมมติฐานศูนย์และสมมติฐานอื่น ๆ สามารถเขียนได้ดังนี้

$$H_0 : SS_{reg} = 0$$

$$H_1 : SS_{reg} > 0$$

สามารถทดสอบนัยสำคัญได้ด้วยสถิติดังนี้

$$F = \frac{[SS_{reg}/m]}{[SS_{res}/(n-m-1)]}$$

เมื่อสัญลักษณ์ที่ใช้ในสูตรเหมือนกับที่ได้อธิบายมาแล้วในสูตรข้างต้น วิธีคิดง่าย ๆ เกี่ยวกับการทดสอบนัยสำคัญทางสถิติด้วย F-test ในรูปแบบทั่ว ๆ ไปคือ

$$F = \frac{[SS_1/df_1]}{[SS_2/df_2]} = MS_1/MS_2$$

นำมาประยุกต์ใช้กับสูตร F-test ทดสอบนัยสำคัญของผลรวมกำลังสอง ได้ว่า

$$F = \frac{[SS_{reg}/df_{reg}]}{[SS_{res}/df_{res}]} = MS_{reg}/MS_{res}$$

เมื่อ $df_{reg} = m$ และ $df_{res} = n – m – 1$ ก่อนหน้านี้สถิติ F-test ที่คำนวณจะนำไปเปรียบเทียบกับค่าวิกฤติ ซึ่งเป็นการทดสอบแบบทางเดียวเสมอ ณ ระดับนัยสำคัญที่กำหนดไว้

กับองศาแห่งความเป็นอิสระ m และ n – m – 1 ซึ่งสามารถหาค่าวิกฤติ F ได้จากตาราง ซึ่งเราสามารถใช้สูตรนี้ทดสอบนัยสำคัญแทนสูตรข้างต้นได้ เพราะว่า

$$r^2 = SS_{reg}/SS_Y$$

และ $$(1 - r^2) = SS_{res}/SS_Y$$

ซึ่ง SS_Y เป็นเทอมที่จะถูกตัดออกจากสมการ ก็จะได้สถิติ F-test เพื่อทดสอบนัยสำคัญดังสูตรข้างต้น

สำหรับตัวอย่างนี้ สามารถแทนค่าสูตรเพื่อคำนวณได้ดังนี้

$$F = \frac{[SS_{reg}/df_{reg}]}{[SS_{res}/df_{res}]} = MS_{reg}/MS_{res}$$

$$= \frac{[427.84/1]}{[80.16/8]} = [427.84]/[10.02]$$

$$= 42.70$$

ค่าวิกฤติที่ระดับนัยสำคัญ .05 จะได้ ${}_{.95}F_{1,8} = 5.32$ ผลของการทดสอบทางสถิติปรากฏว่าค่า F ที่คำนวณได้มากกว่าค่าวิกฤติ F ที่เปิดจากตาราง นั่นคือปฏิเสธ H_0 ยอมรับ H_1 สรุปว่า SS_{reg} มีค่าไม่เท่ากับศูนย์ ที่ระดับนัยสำคัญ .05 (นั่นคือคะแนน GRE สามารถทำนายสัดส่วนของผลรวมกำลังสองของคะแนนสอบกลางภาคได้อย่างมีนัยสำคัญทางสถิติ) ซึ่งสูตรนี้จะได้ค่าเท่ากับสูตรแรก

การทดสอบนัยสำคัญของ b_{YX} (Test of Significance of b_{YX})

การทดสอบที่สามเป็นการทดสอบนัยสำคัญของความชันหรือสัมประสิทธิ์การถดถอย b_{YX} หรือในอีกกรณีหนึ่งก็คือการทดสอบสัมประสิทธิ์การถดถอยที่ไม่เป็นมาตรฐานว่ามีนัยสำคัญแตกต่างจากศูนย์หรือไม่ นั่นคือการทดสอบ b* นั่นเอง เราไม่จำเป็นต้องใช้สูตรแยกกันสำหรับการทดสอบ b* สมมติฐานศูนย์และสมมติฐานอื่น สามารถเขียนได้ดังนี้

$$H_0 : \beta_{YX} = 0$$

$$H_1 : \beta_{YX} \neq 0$$

ในการทดสอบนัยสำคัญส่วนมาก จำเป็นต้องคำนวณความคลาดเคลื่อนมาตรฐานสำหรับการทดสอบสัมประสิทธิ์การถดถอยเท่ากับศูนย์หรือไม่นี้ เราจำเป็นต้องหาความคลาดเคลื่อนมาตรฐานสำหรับ b อย่างไรก็ตาม ก่อนหน้านี้เราได้แนวคิดเกี่ยวกับความคลาดเคลื่อนมาตรฐานของ b แล้ว เราจำเป็นต้องพัฒนาเพิ่มเติมต่อไปอีกในแนวคิดใหม่เกี่ยวกับความคลาดเคลื่อนมาตรฐานดังนี้

แนวคิดใหม่ประการแรกคือ ความแปรปรวนของความคลาดเคลื่อนในการประมาณค่านิยามได้ว่า

$$S^2_{res} = \Sigma e_i^2/df_{res} = SS_{res}/df_{res} = MS_{res}$$

เมื่อผลรวมตั้งแต่ i = 1,…,n สังเกตว่า df_{res} = (n – m – 1) (หรือ n – 2 กรณีตัวแปรทำนายตัวเดียว) องศาแห่งความอิสระไม่ได้มี 2 ค่าเพราะว่าเรามีการประมาณค่าความชันของประชากรและจุดตัด β และ α จากข้อมูลกลุ่มตัวอย่าง ความแปรปรวนของความคลาดเคลื่อนในการประมาณค่าบ่งชี้ปริมาณของความคลาดเคลื่อน ถ้าความคลาดเคลื่อนมีมาก นั่นคือจะมีผลให้ S^2_{res} มีค่าสูง ชี้ให้เห็นถึงความสามารถในการทำนายต่ำ ถ้าความคลาดเคลื่อนมีน้อย จะมีผลให้ค่า S^2_{res} มีค่าต่ำ ชี้ให้ความสามารถในการทำนายสูง

แนวคิดถัดมาคือ ความคลาดเคลื่อนมาตรฐานของการประมาณค่า หรือเรียกว่า root mean square error ความคลาดเคลื่อนมาตรฐานของการประมาณค่า ก็คือรากที่สองของความแปรปรวนของความคลาดเคลื่อนในการประมาณค่าและเรียกได้อีกอย่างว่า ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนในการประมาณค่า ใช้สัญลักษณ์ว่า S_{res}

จำได้ว่าแนวคิดของความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยเกี่ยวข้องกับการแจกแจงปกติ พิสัยของ 1 ส่วนเบี่ยงเบนมาตรฐานที่ห่างจากค่าเฉลี่ย (ค่าเฉลี่ย $\pm$ 1 ส่วนเบี่ยงเบนมาตรฐาน) จะครอบคลุมการประมาณค่า 68 เปอร์เซ็นต์ของการแจกแจง พิสัยของ 2 ส่วนเบี่ยงเบนมาตรฐานที่ห่างจากค่าเฉลี่ย จะครอบคลุม 95 เปอร์เซ็นต์ของการแจกแจง พิสัยของ 3 ส่วนเบี่ยงเบนมาตรฐานที่ห่างจากค่าเฉลี่ย จะครอบคลุม 99 เปอร์เซ็นต์ของการแจกแจง การประยุกต์ใช้แนวคิดนี้ในการทดสอบนัยสำคัญของ b

แนวคิดสุดท้ายคือความคลาดเคลื่อนมาตรฐานของ b เราจะใช้สัญลักษณ์ของความคลาดเคลื่อนมาตรฐานของ b ว่า s_b และนิยามได้ว่า

$$s_b = \frac{s_{res}}{\sqrt{\{[n\Sigma X^2 - (\Sigma X)^2]/n\}}}$$

$$= \frac{s_{res}}{\sqrt{(SS_X)}}$$

เมื่อผลรวมตั้งแต่ i = 1,…,n เราต้องการ s_b ที่น้อยเพื่อจะปฏิเสธ H_0 ดังนั้น s_{res} ควรจะต่ำและ $\sqrt{(SS_X)}$ ควรจะสูง ในกรณีอื่น ๆ เราต้องการให้คะแนน X กระจายมาก ถ้าความแปรปรวนของ X น้อย ก็ไม่แน่ใจว่า X จะสามารถทำนาย Y ได้อย่างมีนัยสำคัญทางสถิติ

จากแนวคิดที่กล่าวมา ในการทดสอบนัยสำคัญทางสถิติของ b มีรูปแบบเป็นอัตราส่วนของการประมาณค่าพารามิเตอร์หารด้วยความคลาดเคลื่อนมาตรฐาน อัตราส่วนของการประมาณค่าพารามิเตอร์ของความชัน b กับความคลาดเคลื่อนมาตรฐาน s_b เขียนได้ในรูป

$$t = b/s_b$$

สถิติทดสอบ t จะเปรียบเทียบกับค่าวิกฤติของ t ในการทดสอบไม่มีทิศทางตาม H_1 ณ ระดับนัยสำคัญที่กำหนดกับองศาแห่งความเป็นอิสระ (n – m – 1) โดยจากตารางค่าวิกฤติ t ที่ $\pm_{(a/2)}t_{(n-m-1)}$ สำหรับการทดสอบแบบไม่มีทิศทาง

สังเกตว่าการทดสอบสถิติ F และ t นั้นจะอธิบายได้ว่า $t^2 = F$

นอกจากนี้ในกรณีที่ข้อมูลเหมือนกัน องศาแห่งความเป็นอิสระเหมือนกัน ระดับนัยสำคัญเดียวกัน การทดสอบทั้ง 3 จะยังคงให้ผลที่เหมือนกัน นั่นคือ ถ้า X สามารถทำนาย Y ได้อย่างมีนัยสำคัญทางสถิติแล้ว H_0 จะถูกปฏิเสธในทั้ง 3 การทดสอบ ถ้า X ทำนาย Y ได้อย่างไม่มีนัยสำคัญทางสถิติ แล้วจะยอมรับ H_0 ทั้ง 3 การทดสอบ ในกรณีของเส้นถดถอยอย่างง่าย ในแต่ละวิธีการทดสอบจะมีสมมติฐานทั่วไปเหมือนกันควรจะนำผู้วิจัยไปสู่ผลการสรุปที่เหมือนกัน

สำหรับสถิติทดสอบที่ 3 นี้ใช้ตัวอย่างคะแนน GRE กับคะแนนสอบกลางภาค จะทดสอบสมมติฐานว่า $H_0 : \beta = 0$ หรือไม่ และเป็นการทดสอบแบบสองทาง ประการแรกเราจะคำนวณความคลาดเคลื่อนของความแปรปรวนในการประมาณค่า ได้

$$s^2_{res} = \Sigma e_i^2/df_{res} = SS_{res}/df_{res} = MS_{res}$$
$$= 80.16/8$$
$$= 10.02$$

ความคลาดเคลื่อนมาตรฐานของการประมาณค่า s_{res}, คำนวณได้ $+\sqrt{10.02} = 3.17$ จากนั้นความคลาดเคลื่อนมาตรฐานของ b จะได้

$$s_b = s_{res} / \sqrt{(SS_X)}$$
$$= 3.17 / \sqrt{(1552.5)}$$
$$= 0.08$$

เมื่อ SS_X คือนำมาจากการคำนวณก่อนหน้านี้ ในที่สุดเราคำนวณสถิติทดสอบ

$$t = b/s_b$$
$$= 0.52/0.08$$
$$= 6.53$$

ประเมินสมมติฐานศูนย์ เราเปรียบเทียบสถิติทดสอบกับค่าวิกฤติ $\pm_{.025}t_8 = \pm 2.306$ สถิติดทสอบมีค่าเกินกว่าค่าวิกฤติ ดังนั้นจะปฏิเสธ H_0 และยอมรับ H_1 สรุปได้ว่าความชันมีนัยสำคัญทางสถิติแตกต่างจากศูนย์ที่ระดับนัยสำคัญ .05 การเปรียบเทียบสถิติ F-test ในการคำนวณสำหรับการทดสอบ 2 อย่างแรกว่ามีนัยสำคัญเช่นเดียวกับการทดสอบ t-test ค่าสถิติ F-test จะคำนวณได้ 42.70 กับ t-test คำนวณ 6.53 เราจะเห็นว่า t^2 จะเท่ากับ F

ช่วงความเชื่อมั่นสำหรับค่าเฉลี่ยที่ถูกทำนายของ Y

วิธีการที่ 4 ในการพัฒนาช่วงความเชื่อมั่นสำหรับค่าเฉลี่ยที่ถูกทำนายของ Y เราใช้สัญลักษณ์ว่า $\overline{Y}'_0$ ณ ค่าของ X_0 ในอีกกรณีหนึ่ง $\overline{Y}'_0$ จะถูกอ้างอิงว่าเป็นค่าคาดหวังของ Y หรือค่าเฉลี่ยที่มีเงื่อนไขของ Y เมื่อกำหนดค่า X ในอีกกรณีหนึ่งค่าคะแนนทำนายเฉพาะของ X_0 สามารถคำนวณช่วงความเชื่อมั่นของค่าเฉลี่ยที่ถูกทำนายของ Y ได้ดังนี้

ความคลาดเคลื่อนมาตรฐานของ $\overline{Y}'_0$ คือ

$$s(\overline{Y}'_0) = s_{res}\sqrt{\{(1/n) + [(X_0 - \overline{X})^2 / SS_x]\}}$$

ความคลาดเคลื่อนมาตรฐานขึ้นอยู่กับค่าเฉพาะของ X_0 ที่เลือก เราคาดหวังว่าความสามารถในการทำนายที่น้อยที่สุดคือค่าที่อยู่ห่างจากค่า X มากที่สุด ดังนั้นค่าของตัวทำนายที่อยู่ใกล้ศูนย์กลางของการแจกแจงของคะแนน X มากที่สุด จะมีความสามารถในการทำนายสูงสุด

ช่วงความสามารถของ $\overline{Y}'_0$ มีรูปแบบดังนี้

$$CI(\overline{Y}'_0) = \overline{Y}'_0 \pm_{(\alpha/2)} t_{(n-2)} s(\overline{Y}'_0)$$

ให้เราพิจารณาตัวอย่างช่วงความเชื่อมั่นกับข้อมูลตัวอย่าง ถ้าเรานำ GRE คะแนน 50 มาทำนายคะแนนสอบกลางภาคควรจะได้ 35.11 ช่วงความเชื่อมั่นสำหรับการคะแนนของ 35.11 คำนวณได้ดังนี้

$$\begin{aligned} s(\overline{Y}'_0) &= s_{res}\sqrt{\{(1/n) + [(X_0 - \overline{X})^2 / SS_x]\}} \\ &= \sqrt{3.17\{(1/10) + [(50-55)^2 / 1552.5]\}} \\ &= 1.08 \end{aligned}$$

และ

$$\begin{aligned} CI(\overline{Y}'_0) &= \overline{Y}'_0 \pm_{(\alpha/2)} t_{(n-2)} s(\overline{Y}'_0) \\ &= \overline{Y}'_0 \pm_{.025} t_8 s(\overline{Y}'_0) \\ &= 35.11 \pm (2.306)(1.08) \\ &= 35.11 \pm 2.49 \\ &= (32.62, 37.60) \end{aligned}$$

กลับไปยังภาพประกอบ 3 ช่วงความเชื่อมั่นของ $\overline{Y}'_0$ ที่ค่า X_0 จะถูกพล็อตเป็นคู่ของจุดที่ใกล้เส้นถดถอยมากที่สุด

ช่วงการทำนายสำหรับค่าของ Y

กระบวนการที่ห้าและกระบวนการสุดท้ายสำหรับช่วงของการทดสอบที่ค่าเฉลี่ยของ Y ค่าของ Y'_0 ที่ค่าเฉพาะของ X_0 นั่นคือคะแนนทำนายสำหรับค่าเฉลี่ยที่ทราบค่า แต่คะแนนเกณฑ์สำหรับค่าเฉลี่ยยังไม่ถูกสังเกต นั่นคือในการเปรียบเทียบช่วงความเชื่อมั่นจะพิจารณาเมื่อคะแนนเฉพาะ Y พร้อมที่จะสังเกตได้ ดังนั้นช่วงความเชื่อมั่นกับค่าเฉลี่ยของค่าที่ถูกทำนายเมื่อช่วงของการทำนายกับค่าที่ถูกทำนายเฉพาะคนยังไม่ได้ถูกสังเกต

ความคลาดเคลื่อนมาตรฐานของ Y'_0 คือ

$$s(Y'_0) = s_{res}\sqrt{\{1 + (1/n) + [(X_0 - \overline{X})^2 / SS_x]\}}$$

ความคลาดเคลื่อนมาตรฐานของ Y'_0 คล้ายกับความคลาดเคลื่อนมาตรฐานของ $\overline{Y}'_0$ กับเพิ่ม s_{res} ในสมการ ดังนั้นความคลาดเคลื่อนมาตรฐานของ Y'_0 จะมีค่ามากกว่าความคลาดเคลื่อน

มาตรฐานของ $\overline{Y}'_0$ เสมอ ดังนั้นความคลาดเคลื่อนมาตรฐานขึ้นอยู่กับค่าเฉพาะของ X ที่เลือก ซึ่งเราจะมีความเชื่อมั่นได้มากในการทำนายสำหรับค่าของ X ที่เข้าใกล้ $\overline{X}$

ช่วงการทำนาย (PI) ของ Y'_0 มีรูปแบบดังนี้

$$PI(Y'_0) = Y'_0 \pm {}_{(a/2)}t_{(n-2)}s(Y'_0)$$

ดังนั้นช่วงการทำนายจะมีความกว้างมากกว่าเสมอ

เรามาดูตัวอย่างในการใช้กระบวนการช่วงการทำนายกับตัวอย่าง ถ้าเรานำคะแนน GRE ที่ 50 มาทำนายคะแนนสอบกลางภาคจะได้ 35.11 ช่วงการทำนายสำหรับค่าที่ถูกทำนายเฉพาะที่ 35.11 ควรจะได้ว่า

$$s(Y'_0) = s_{res}\sqrt{\{1 + (1/n) + [(X_0 - \overline{X})^2 / SS_x]\}}$$

$$= \sqrt{3.17\{1 + (1/10) + [(X_0 - \overline{X})^2 / SS_x]\}}$$

$$= 3.34$$

และ

$$PI(Y'_0) = Y'_0 \pm {}_{(a/2)}t_{(n-2)}s(Y'_0)$$

$$= Y'_0 \pm {}_{.025}t_8 s(Y'_0)$$

$$= 35.11 \pm (2.306)(3.34)$$

$$= 35.11 \pm 7.71$$

$$= (27.40, 42.82)$$

ในภาพประกอบ 3 ช่วงการทำนายของ Y'_0 เมื่อให้ค่า X_0 สามารถพล็อตเป็นคู่ของจุดบนเส้นที่อยู่ไกลเส้นถดถอยที่สุด ดังนั้นช่วงการทำนายจะมีพิสัยกว้างกว่าช่วงความเชื่อมั่นเสมอ

บรรณานุกรม

Darlington, Richard B. **Regression and Linear Models**. USA : McGraw-Hill Publishing Company, 1990.

Lomax, Richard G. **Statistical Concepts : A Second Course for Education and the Behavioral Sciences**. London : Lawrence Erlbaum Associates, Inc., 1992.