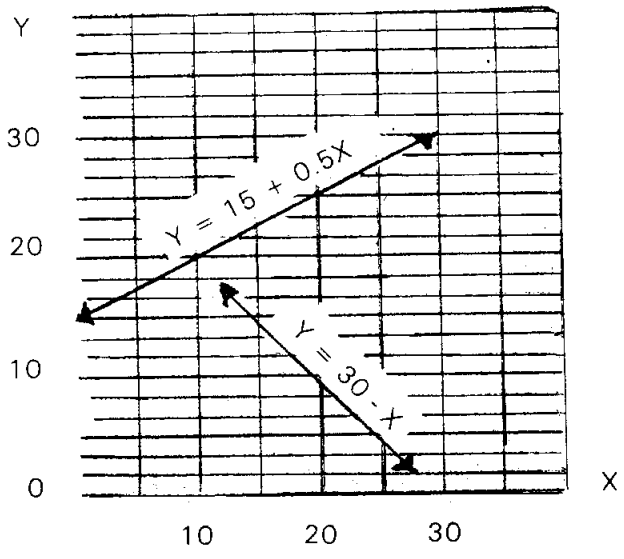
# 11. ความถดถอยเชิงเดี่ยวและสหสัมพันธ์

1. บทนำ
2. การประมาณค่าโดยใช้เส้นถดถอย
3. การวิเคราะห์สหสัมพันธ์
4. แบบฝึกหัด

## 1. บทนำ

เมื่อเราวิเคราะห์ข้อมูลที่มีตัวแปรเพียงตัวเดียว เราเรียกว่า การวิเคราะห์**ตัวแปรเชิงเดี่ยว** หรือ Univariate analysis แต่ถ้าปัญหาของเรามีตัวแปร 2 ตัว และเราต้องการทราบความสัมพันธ์ของตัวแปรคู่นั้น เช่น ผู้จัดการฝ่ายขายอาจสนใจใคร่ทราบความสัมพันธ์ระหว่างจำนวนรายได้จากการขายกับค่าใช้จ่ายในการโฆษณา ฝ่ายทะเบียนของวิทยาลัยอาจสนใจความสัมพันธ์ระหว่างคะแนนในระดับมัธยมปลายกับคะแนนในระดับวิทยาลัย ผู้จัดการฝ่ายบุคคลอาจสนใจความสัมพันธ์ของคะแนนทดสอบความถนัดกับผลงานของลูกจ้าง นักเกษตรอาจสนใจความสัมพันธ์ระหว่างจำนวนปุ๋ยและผลผลิต นอกจากเราต้องการทราบความสัมพันธ์ระหว่างตัวแปรแต่ละคู่แล้ว บางครั้งเรายังต้องการพยากรณ์ค่าของตัวแปรตัวหนึ่งโดยใช้ค่าของตัวแปรอีกตัวหนึ่งด้วย และเราเรียกวิธีการนี้ว่า การศึกษาความถดถอยและสหสัมพันธ์

การศึกษาความถอดถอย จะหมายถึงการหาสมการที่สอดคล้อง (fit) กับข้อมูลที่เก็บมา เพื่อใช้สมการนั้น **พยากรณ์** (prediction) และประมาณค่า (estimation) ตัวแปรที่เราพยากรณ์ได้หรือประมาณค่าได้เรียกว่า **ตัวแปรพึ่งพา** (dependent variable) และตัวแปรอีกตัวหนึ่งซึ่งเป็นพื้นฐานที่ใช้ประมาณค่า เรียกว่า **ตัวแปรอิสระ** (independent variable) **การศึกษาความถดถอยแบบเชิงเดี่ยว** (simple regression) คือการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระ 1 ตัว กับตัวแปรพึ่งพา 1 ตัว ส่วน**การศึกษาความถดถอยแบบเชิงซ้อน** (multiple regression) จะมีตัวแปรอิสระ 2 ตัวขึ้นไป และตัวแปรพึ่งพา 1 ตัว ในบทนี้ จะศึกษาการสร้างสมการถดถอยเชิงเดี่ยว ซึ่งมี X เป็นตัวแปรอิสระ และ Y เป็นตัวแปรพึ่งพา และเรานิยมเรียกว่า**ความถดถอยของ** Y **บน** X (regression of Y on X)




**รูปที่** 11.1
แสดงสมการเส้นตรง 2 เส้น

เราจะจำกัดความสนใจศึกษาเฉพาะ**ความถดถอยเชิงเดี่ยว**เท่านั้น นั่นคือเราจะสามารถแสดงความสัมพันธ์โดยกราฟในรูปของเส้นตรง และมีอยู่บ่อยครั้งที่สมการถดถอยเป็นแบบเส้นโค้ง (curvilinear) ในรูปที่ 11.1 จะมีเส้นถดถอย 2 เส้น เส้นหนึ่งมีความลาดชันแบบดิ่งขึ้น (upward sloping) คือเส้นที่แสดงโดยสมการ

$$Y = 15 + 0.5 X$$

ส่วนเส้นที่มีความลาดชันแบบดิ่งลง (downward sloping) คือ เส้น

$$Y = 30 - X$$

ดังนั้น รูปแบบของสมการถดถอยเชิงเดี่ยว คือ

$$Y = a + bX$$

ในเมื่อ

a คือ จุดที่เส้นถดถอยตัดแกน Y (Y intercept)

b คือ ความลาดชันของเส้นตรง หมายถึง อัตราการเปลี่ยนแปลงของ Y ต่อทุก ๆ หน่วยของ X ที่เปลี่ยนแปลง

ดังนั้น เราจะสร้างสมการถดถอยได้ก็เมื่อทราบค่า a และ b การที่เราพยายามสร้างสมการแสดงความสัมพันธ์ระหว่างตัวแปรคู่หนึ่งก็เพื่อใช้ประโยชน์ในการพยากรณ์ โดยใช้ค่าของตัวแปรตัวหนึ่งทำนายค่าของตัวแปรอีกตัวหนึ่ง เช่น ถ้าเราทราบความสัมพันธ์ระหว่างคะแนนความถนัดกับ GPA เราก็สามารถทำนายผลการเรียน หรือ GPA จากคะแนนความถนัด ในทำนองเดียวกัน ถ้าเราทราบความสัมพันธ์ระหว่างเงินกู้เพื่อปลูกสร้างที่อยู่อาศัยกับจำนวนบ้านจัดสรร เราก็จะสามารถพยากรณ์จำนวนบ้านจัดสรรที่จะก่อสร้างในแต่ละปีได้จากงบประมาณเงินกู้เพื่อสร้างที่อยู่อาศัย

ความสัมพันธ์ระหว่างตัวแปรคู่หนึ่งนั้น ไม่จำเป็นต้องอยู่ในรูปความสัมพันธ์แบบ**เหตุและผล**เสมอไป (cause-and-effect หรือ causal relationships) ซึ่งหมายถึงตัวแปรอิสระเป็นต้นเหตุของตัวแปรพึ่งพา ในบางครั้งความสัมพันธ์แบบนี้เป็นจริง เช่น ผลผลิตข้าวกับปุ๋ย ปุ๋ยเป็นสาเหตุที่ทำให้ผลผลิตสูง แต่ความสัมพันธ์บางคู่ เช่น คะแนนความถนัดกับ GPA ซึ่งเราทราบว่าน่าจะมีความสัมพันธ์แบบเชิงบวก แต่เราคงจะยอมรับไม่ได้ว่าคะแนนความถนัดเป็นสาเหตุของ GPA ทั้งนี้ เป็นเพราะยังมีปัจจัยอื่น ๆ อีกหลายอย่างที่มีอิทธิพลต่อ GPA จึงสรุปได้ว่า ความสัมพันธ์ระหว่างตัวแปรบางคู่อยู่ในลักษณะเหตุและผล แต่หลายคู่ที่อยู่ในลักษณะฟังก์ชั่นหรือผลที่ตามมา (functional or consequential) คือ มีแฟคเตอร์ที่สามเข้ามาเกี่ยวข้องด้วย

เมื่อเราสร้างสมการถดถอยเพื่อแสดงความสัมพันธ์ระหว่างตัวแปรคู่หนึ่งแล้ว เราอาจต้องการทราบว่า ตัวแปรคู่นั้น มีความสัมพันธ์กันมากน้อยเพียงใดซึ่งจะต้องอาศัยเทคนิคของ**การวิเคราะห์สหสัมพันธ์** (correlation analysis) ซึ่งจะวัดความใกล้ชิดของความสัมพันธ์เชิงเส้นระหว่างตัวแปร 2 ตัวในสมการถดถอย แต่บางครั้งเราอาจต้องการวัดความใกล้ชิดของความสัมพันธ์ระหว่างตัวแปรคู่หนึ่งโดยยังไม่ได้ศึกษาความถดถอยก็ได้ นั่นคือ เราสามารถจะศึกษาความถดถอย และการวิเคราะห์สหสัมพันธ์โดยอิสระต่อกัน

---

## แบบฝึกหัด

11.1 จงบอกลักษณะของสมการข้างล่างว่าอันใด เป็นแบบเชิงเดี่ยว, เชิงซ้อน, เส้นตรง และเส้นโค้ง

ก) $Y = 2 + 3X$

ข) $Y = 1000 - 5000X$

ค) $Y = 10 + 2X_1 + 3X_2 + 4X_3$

ง) $Y = 2 + 3X + X^2$

11.2 จงบอกจุดประสงค์ของการวิเคราะห์ความถดถอยและการวิเคราะห์สหสัมพันธ์

11.3 จงอธิบายความหมาย "ความถดถอย Y บน X" ตัวใดเป็นตัวแปรอิสระ และตัวใดเป็นตัวแปรพึ่งพา

11.4 จากสมการ $Y = a + bX$ จงอธิบายความหมายของค่า a และ b โดยรูปกราฟ

11.5 จงกราฟสมการ $Y = a + bX$ เมื่อ a และ b มีค่าต่าง ๆ ดังนี้

ก) $a = 2, b = 0$

ข) $a = 0, b = 1$

ค) $a = 2, b = 1$

ง) $a = -2, b = 2$

จ) $a = 4, b = -2$

---

## 2. ความถดถอยเชิงเส้น

(linear regression)

เมื่อเก็บข้อมูลได้แล้วควรพล็อตข้อมูลทั้งหมดในกราฟ เรียกว่า scatter diagram เพื่อให้เรามองดูด้วยตาเปล่าก่อนว่าลักษณะความสัมพันธ์ระหว่างตัวแปร X และ Y มีลักษณะแบบเส้น

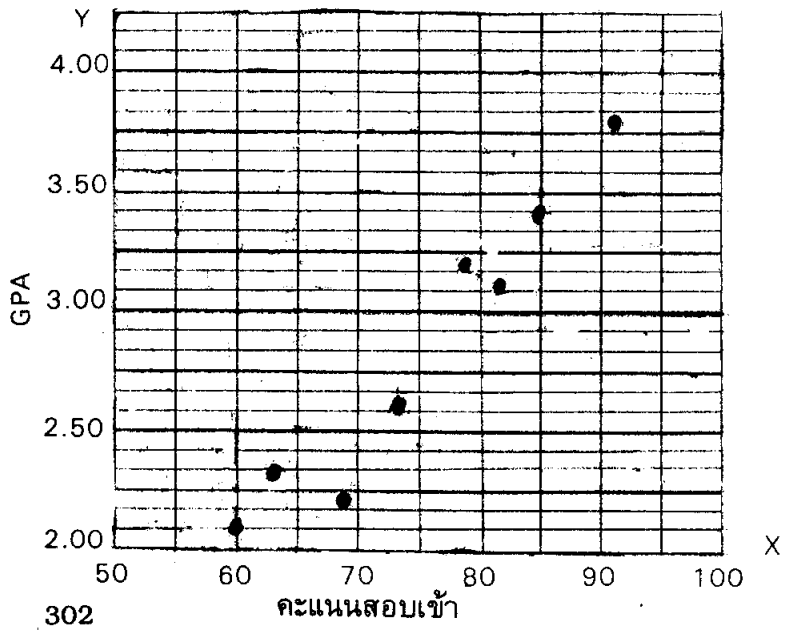
โค้ง หรือเส้นตรงโดยเราจะใช้แกน X แทนค่าต่าง ๆ ของตัวแปร X และแกน Y แทนค่าต่าง ๆ ของตัวแปร Y ข้อมูลที่เก็บมาจะเป็นคู่ลำดับของ X และ Y n คู่ คือ $(x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n)$ เราจึงใช้คู่ลำดับเหล่านี้คือจุดต่าง ๆ n จุดนั่นเอง กล่าวโดยสรุปได้ว่าประโยชน์ของ scatter diagram มี 2 อย่าง คือ 1. ใช้เป็นข้อบ่งชี้ว่า ตัวแปรคู่นั้นมีความสัมพันธ์กันหรือไม่ 2. ถ้ามีความสัมพันธ์กัน เราจะดูต่อไปว่าจะแทนความสัมพันธ์นั้นด้วยเส้นแบบใด นั่นคือ จะหาสมการประมาณค่าความสัมพันธ์นั้น

**ตัวอย่าง** ต้องการศึกษาว่า คะแนนสอบเข้ามหาวิทยาลัยและ GPA มีความสัมพันธ์กันหรือไม่ ฝ่ายทะเบียนได้เก็บข้อมูลไว้ในตารางที่ 11.1 ดังนี้

**ตารางที่** 11.1 คะแนนสอบเข้า และ GPA เมื่อจบการศึกษาของนักศึกษา 8 คน

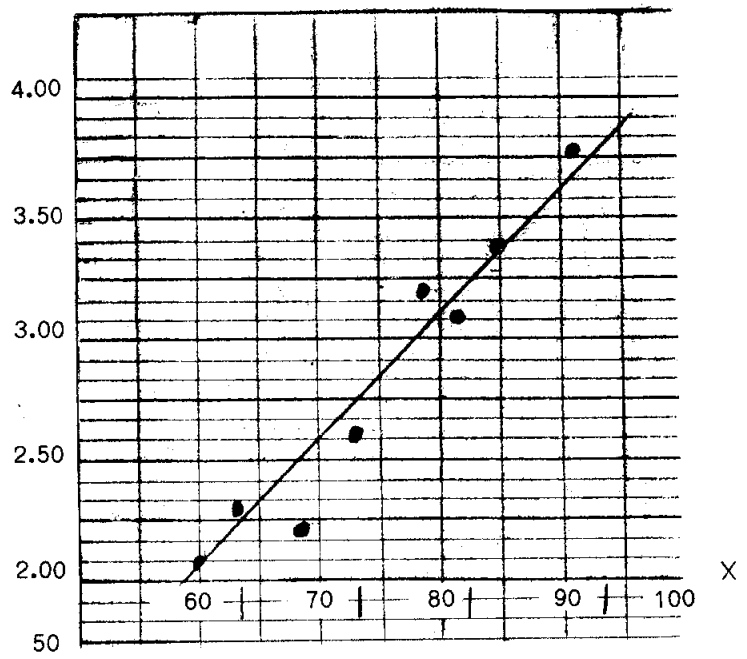
| นักศึกษา | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| คะแนนสอบเข้า (เต็ม 100 คะแนน) | 74 | 69 | 85 | 63 | 82 | 60 | 79 | 91 |
| GPA (4.0 = A) | 2.6 | 2.2 | 3.4 | 2.3 | 3.1 | 2.1 | 3.2 | 3.8 |

ในขั้นแรก เราควรย้ายข้อมูลในตาราง 11.1 ลงในกราฟก่อน และเนื่องจากฝ่ายทะเบียนต้องการใช้คะแนนสอบเข้าเป็นตัวพยากรณ์ความสำเร็จของนักศึกษาในขั้นมหาวิทยาลัย จึงควรให้ GPA เป็นตัวแปรพึ่งพาและแทนด้วยแกน Y ส่วนคะแนนสอบเข้าให้เป็นตัวแปรอิสระและแทนด้วยแกน X ดังแสดงในรูปที่ 11.2



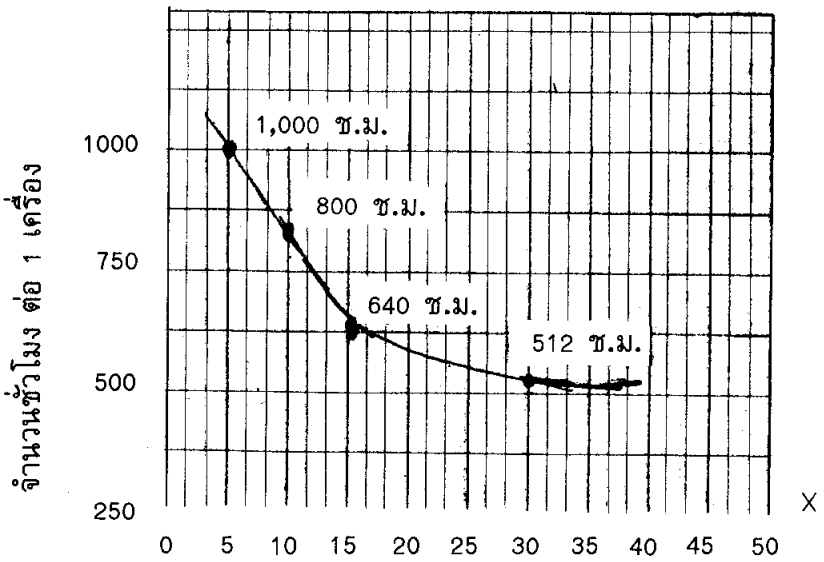

**รูปที่** 11.2
scatter diagram
ของข้อมูลในตาราง 11.1

เมื่อพิจารณารูปที่ 11.2 ด้วยตาเปล่าจะพบว่า จุด 8 จุด ซึ่งคือคู่ลำดับ $(x_i, y_i)$ 8 คู่ แสดงว่าตัวแปรคู่นี้น่าจะมีความสัมพันธ์กัน และหากว่าเราลากเส้นตรงเส้นหนึ่งให้สอดไประหว่างกึ่งกลางของกลุ่มของจุดเหล่านี้ โดยให้จำนวนจุดที่อยู่เหนือและใต้เส้นมีจำนวนเท่า ๆ กัน ดังรูปที่ 11.3 โดยเส้นตรงนั้นจะแสดงความสัมพันธ์โดยตรงของ X และ Y แต่เนื่องจากทุกจุดมิได้อยู่บนเส้นตรงทั้งหมด จึงกล่าวได้ว่า คะแนนสอบเข้า (X) และ GPA (Y) มีความโน้มเอียงค่อนข้างสูงที่จะมีความสัมพันธ์กันเชิงเส้นตรง (linear relationship)




**รูปที่** 11.3
scatter digram และเส้นตรงแสดงความสัมพันธ์ระหว่าง X และ Y

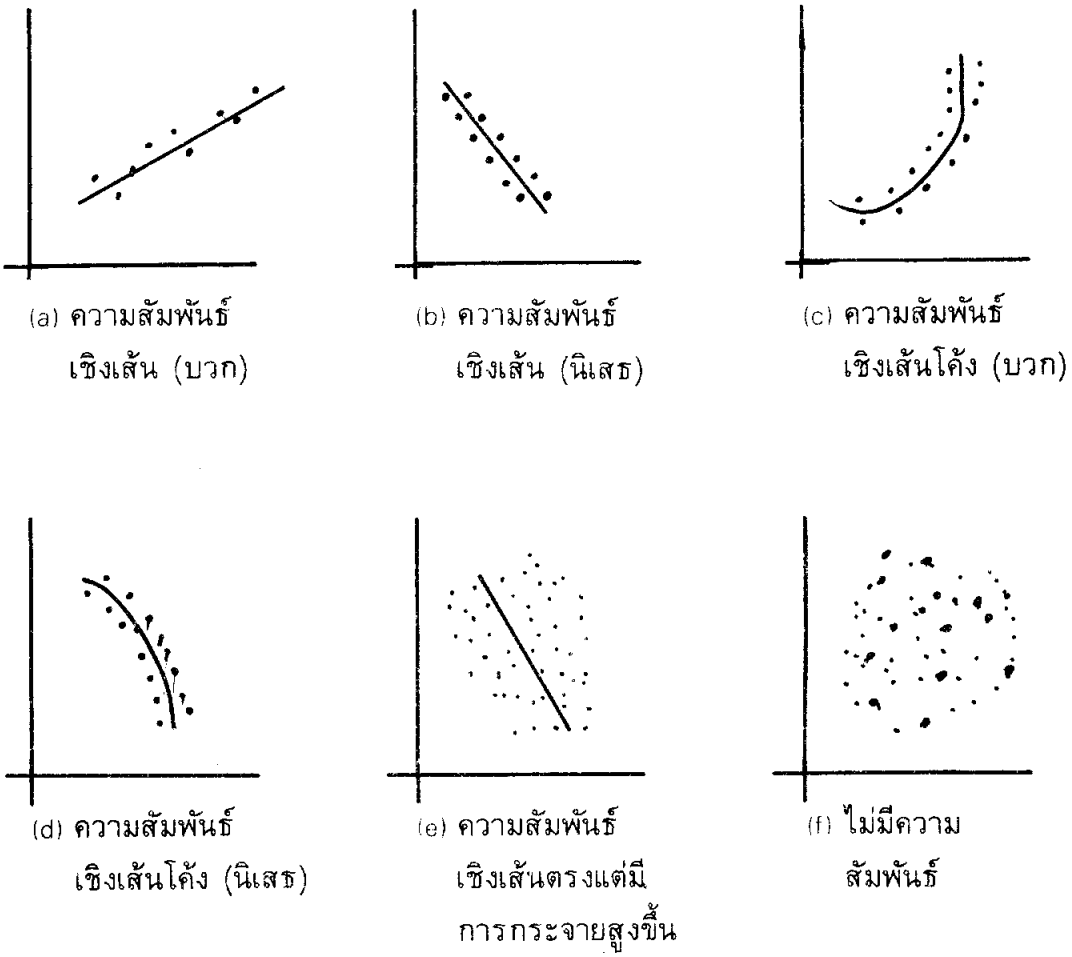
บางครั้งความสัมพันธ์ระหว่าง X และ Y อาจเป็นรูปโค้ง เราเรียกว่ามีความสัมพันธ์แบบเส้นโค้ง เช่นความสัมพันธ์ระหว่างทักษะกับเวลาที่ใช้ในการผลิต ได้มีผู้เก็บสถิติจากพนักงานโรงงานประกอบเครื่องบิน พบว่า พนักงานที่มีทักษะสูงจะใช้เวลาประกอบเครื่องบินลดลง 20% ทุกครั้งที่จำนวนผลิตเพิ่มขึ้น 1 เท่าตัว ดังรูปที่ 11.4 เมื่อผลิตเครื่องบิน 5 ลำ ใช้เวลา 1000 ชั่วโมง ต่อ 1 เครื่อง แต่ระหว่างที่ 6-10 จะใช้เวลาผลิตต่อเครื่องลดลงเหลือ 800 ชั่วโมง (ลดลง 20%) เครื่องที่ 11-15 ใช้เวลาลดลงอีก 20% เหลือ 640 ชั่วโมงต่อเครื่อง




**รูปที่** 11.4

แสดงความสัมพันธ์แบบเส้นโค้งระหว่าง เวลาที่ใช้ผลิต และจำนวนเครื่องที่ผลิตแล้ว บางครั้งเรียกว่าโค้งการเรียนรู้ (learnig curve)

เราจะทราบว่าตัวแปรมีความสัมพันธ์แบบตามกันเรียกว่า**เชิงบวก** หรือมีความสัมพันธ์ในทิศทางตรงข้ามเรียกว่า **เชิงนิเสธ** โดยดูทิศทางของเส้นตรงหรือเส้นโค้ง ในรูป 11.3 ความสัมพันธ์เป็นแบบเชิงบวก ในรูป 11.4 ความสัมพันธ์เป็นแบบนิเสธ และในรูป 11.5 จะแสดงความสัมพันธ์แบบต่าง ๆ รูป a และ b เป็นความสัมพันธ์เชิงบวกและเชิงนิเสธ รูป c และ d เป็นความสัมพันธ์




**รูปที่** 11.5
แสดงความสัมพันธ์
แบบต่าง ๆ ระหว่าง
X และ Y ใน
scatter diagram

แบบเส้นโค้งเชิงบวกและเชิงนิเสธ รูป e แสดงความสัมพันธ์เชิงนิเสธ และมีการกระจายของจุดโดยรอบเส้นตรงค่อนข้างสูงแสดงว่าดีกรีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระและตัวแปรพึ่งพาจะมีขนาดเล็กกว่าความสัมพันธ์ในรูป b ส่วนลักษณะการเรียงตัวของจุดในรูป f ไม่มีรูปแบบเด่นชัด จึงแสดงว่า ไม่มีความสัมพันธ์ระหว่างตัวแปรคู่นั้น ดังนั้น จึงไม่สามารถใช้ข่าวสารของตัวแปรหนึ่งพยากรณ์การเกิดขึ้นของตัวแปรอีกตัวหนึ่ง

ปัญหาคือ เส้นตรงนั้นมาจากไหน วิธีง่ายที่สุดแต่ค่อนข้างหยาบคือการลากเส้นตรงโดย**วิธีกะเอา** (freehand method) แต่มีข้อเสีย คือ เราไม่ทราบว่าเป็นเส้นที่ปรับเข้ากับข้อมูล (fit) ได้ดีที่สุดหรือไม่

## วิธีกำลังสองน้อยที่สุด

(The Method of Least Squares)

เราสามารถลากเส้นผ่านจุดเหล่านั้นได้หลายเส้นเป็นจำนวนนับไม่ถ้วน มีหลายเส้นที่ไม่สามารถปรับเข้ากับข้อมูลได้ดี จึงควรตัดทิ้งไปแต่ก็ยังคงเหลืออีกหลายเส้นที่น่าจะปรับเข้ากับข้อมูลได้ดี แต่เราต้องการเพียง**เส้นเดียวที่ปรับเข้ากับข้อมูลได้ดีที่สุด** จึงควรต้องนิยามคำว่า **"ดีที่สุด"** ถ้าทุก ๆ จุดอยู่บนเส้นตรงทั้งหมด เราจะเห็นชัดเจนว่าเส้นนั้นปรับเข้ากับข้อมูลได้ดีที่สุด แต่ลักษณะเช่นนี้ จะไม่ค่อยปรากฏ ส่วนใหญ่จะเป็นลักษณะจุดเหล่านั้นกระจายอยู่โดยรอบเส้นตรง อาจมีบางจุดอยู่บนเส้นตรง แต่ไม่ใช่ทั้งหมด เราจึงควรขยายความว่า เส้นที่ดีปรับเข้ากับข้อมูลได้ดีที่สุดจะต้องมีคุณสมบัติอะไรบ้าง หลักเกณฑ์ที่นิยมใช้ทั่วไปคือ **วิธีกำลังสองน้อยที่สุด** (least squares method) ซึ่งหมายถึงเส้นที่ให้ผลรวมกำลังสองของระยะทางแนวดิ่งระหว่างจุดกับเส้นตรงของทุก ๆ จุด มีค่าน้อยที่สุด

เมื่อนำข้อมูลในตารางที่ 11.2 พล็อตใน Scatter diagram และลากเส้นตรง $Y = a + bX$ ผ่านระหว่างกลางของจุดเหล่านี้ ในรูปที่ 11.6 จุด 20 จุดนั้นคือ $(X_1, Y_1)$ , $(X_2, Y_2)$ ,... $(X_{20}, Y_{20})$ ส่วนสมการเส้นตรงที่ปรับเข้ากับจุด 20 นี้ คือ

$$Y = a + bX$$

ดังนั้น ทุก ๆ ค่าของ X จะมีค่า Y จำนวน 2 ค่า คือข้อมูลที่เก็บมาซึ่งคู่กับค่า X ที่เก็บมาในตารางที่ 11.2 กับอีกค่าหนึ่งคือค่า Y ที่อยู่บนเส้นตรง ซึ่งได้จากการแทนค่า X ลงในสมการ $Y = a + bX$ เช่นที่ $X_1 = 3$ ค่า $Y_1 = 5$ อันนี้คือ ค่า observed Y ยังมีค่า $Y_1$ อีกอันหนึ่งคือ $Y_1 = a + bX_1$ และจุด $(X_1, Y_1)$ จะอยู่บนเส้นตรง และเป็นเช่นนี้จนถึง $X_{20}$ จะมี obseved $Y_{20} = 5$ และ $Y_{20} = a + bX_{20}$ ความแตกต่างของค่า Y คู่นี้ สำหรับแต่ละค่าของ X คือระยะทางแนวดิ่งในรูป 11.6 และถ้าผลต่างกำลังสองของทุก ๆ คู่ลำดับคือ

$$\Sigma_{i=1}^{n} \quad (Y_i - (a + bX_i) \quad )^2$$

เป็นค่าน้อยที่สุด เราจะเรียกเส้นนั้นว่า **เส้นกำลังสองน้อยที่สุด** ค่า a คือ จุดที่เส้นตรงตัดแกน Y (Y intercept) และ b คือความลาดชันของเส้น เราเรียก a และ b ว่า**สัมประสิทธิ์ความถดถอย** (regression coeffients) ในขณะนี้สมมุติเรามีข้อมูลคือ คะแนนผลงานและ GPA ของพนักงาน 20 คน ในตารางที่ 11.2 โดยให้ GPA เป็นตัวแปรอิสระ (X) และคะแนนผลงานเป็นตัวแปรพึ่งพา (Y) เพราะต้องการพยากรณ์คะแนนผลงาน จาก GPA

**ตารางที่** 11.2 คะแนนผลงานและ GPA เมื่อจบการศึกษาของพนักงาน 20 คน

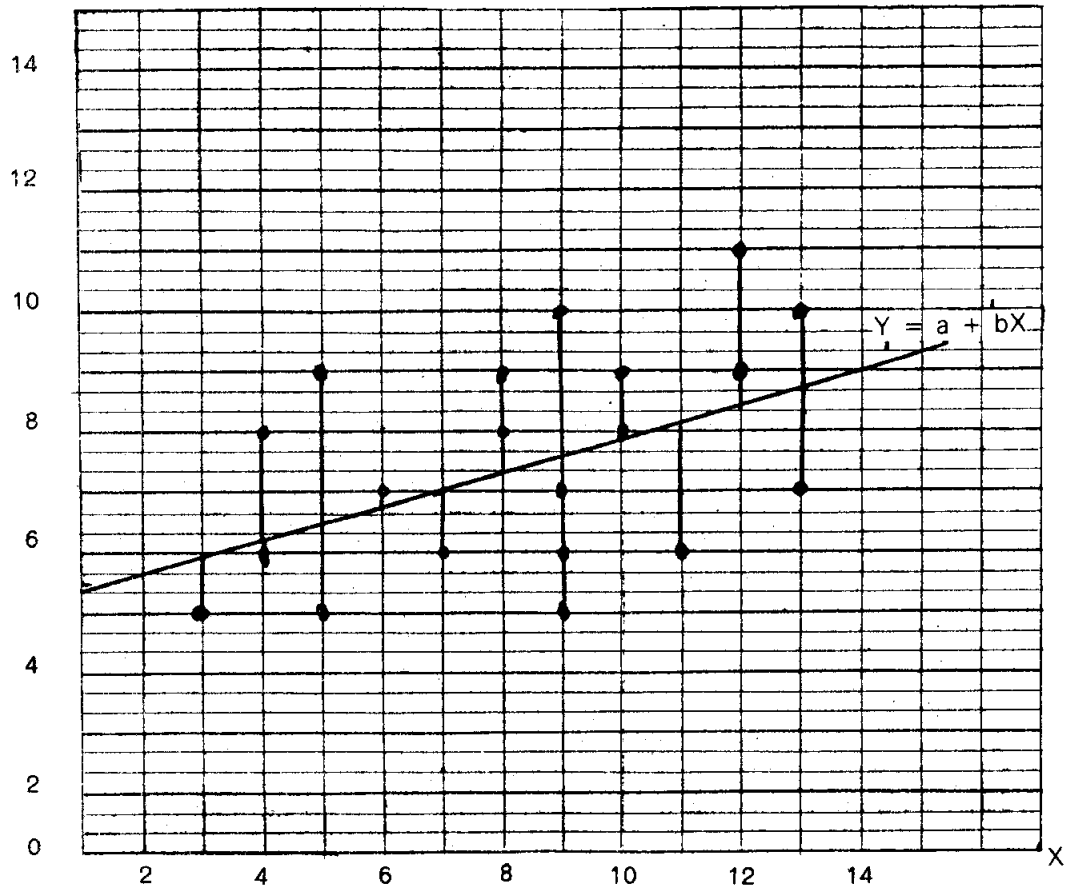
| พนักงาน | GPA เมื่อจบการศึกษา (X) | คะแนนผลงาน (Y) |
|---|---|---|
| 1 | 3 | 5 |
| 2 | 2 | 4 |
| 3 | 4 | 4 |
| 4 | 12 | 9 |
| 5 | 11 | 8 |
| 6 | 8 | 9 |
| 7 | 9 | 7 |
| 8 | 7 | 8 |
| 9 | 6 | 5 |
| 10 | 5 | 6 |
| 11 | 4 | 8 |
| 12 | 8 | 4 |
| 13 | 3 | 7 |
| 14 | 12 | 6 |
| 15 | 9 | 8 |
| 16 | 8 | 5 |
| 17 | 11 | 10 |
| 18 | 7 | 7 |
| 19 | 8 | 6 |
| 20 | 10 | 5 |

จะเห็นว่า เราจะสร้างเส้นตรงทั้งหลายได้ก็ต่อเมื่อทราบค่า a และ b เราจึงต้องการหาค่า a และ b ที่ทำให้เส้นตรงนั้นปรับเข้ากับข้อมูลได้ดีที่สุดตามเกณฑ์กำลังสองน้อยที่สุด ซึ่งโดยวิธีคณิตศาสตร์เราจะหาค่า a และ b ได้จากสมการปกติ (normal equations) คือ สมการที่ (11.1) และ (11.2)

$$\Sigma Y = na + b\Sigma X \qquad (11.1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad (11.2)$$

ในเมื่อ n คือจำนวนคู่ลำดับ (X, Y) ส่วนค่าอื่น ๆ นอกจากค่า a และ b จะหาได้จากข้อมูลในตารางที่ 11.2 เราจึงหาค่า a และ b จากสมการปกติ




**รูปที่** 11.6
แสดงระยะแนวดิ่งจากจุดต่าง ๆ
กับเส้นตรงของข้อมูลในตารางที่ 11.2

**ตารางที่** 11.3 แสดงการคำนวณเพื่อหาความถดถอยของคะแนนผลงานและ GPA เมื่อจบการศึกษา

| พนักงาน | GPA (X) | (Y) | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 15 | 9 | 25 |
| 2 | 2 | 4 | 8 | 4 | 16 |
| 3 | 4 | 4 | 16 | 16 | 16 |
| 4 | 12 | 9 | 108 | 144 | 81 |
| 5 | 11 | 8 | 88 | 121 | 64 |
| 6 | 8 | 9 | 72 | 64 | 81 |
| 7 | 9 | 7 | 63 | 81 | 49 |
| 8 | 7 | 8 | 56 | 49 | 64 |
| 9 | 6 | 5 | 30 | 26 | 25 |
| 10 | 5 | 6 | 30 | 25 | 36 |
| 11 | 4 | 8 | 32 | 16 | 64 |
| 12 | 8 | 4 | 32 | 64 | 16 |
| 13 | 3 | 7 | 21 | 9 | 49 |
| 14 | 12 | 6 | 72 | 144 | 36 |
| 15 | 9 | 8 | 72 | 81 | 64 |
| 16 | 8 | 5 | 40 | 64 | 25 |
| 17 | 11 | 10 | 110 | 121 | 100 |
| 18 | 7 | 49 | 7 | 49 | 49 |
| 19 | 8 | 6 | 48 | 64 | 36 |
| 20 | 10 | 5 | 50 | 100 | 25 |
| n = 20 | 147 = $\Sigma X$ | 131 = $\Sigma Y$ | 1012 = $\Sigma XY$ | 1261 = $\Sigma X^2$ | 921 = $\Sigma Y^2$ |

เมื่อใช้ค่าที่คำนวณได้ในตารางที่ 11.3 แทนค่าในสมการปกติ จะได้

(1) $131 = 20\,a + 147\,b$

(2) $1012 = 147\,a + 1261\,b$

เอา 147 คูณสมการ (1) และ 20 คูณสมการ (2) แล้วเอาสมการที่ (2) หักออกจากสมการที่ (1)

(1) $19{,}257 = 2940\,a + 21{,}609\,b$

(2) $20{,}240 = 2940\,a + 25{,}220\,b$

เมื่อลบกันแล้ว จะได้

$$983 = 3611\,b$$

ดังนั้น

$$b = \frac{983}{3611} = 0.2722$$

แทนค่า b ในสมการ (1) จะได้

$$131 = 20\,a + 147(0.2722)$$

$$20\,a = 131 - 40.0134 = 90.9866$$

ดังนั้น

$$a = \frac{90.9866}{20} = 4.55$$

และสมการกำลังสองน้อยที่สุด คือ

$$Y = 4.55 + 0.27\,X$$

ในปัจจุบันนี้ เราใช้เครื่องคิดเลขขนาดเล็กอย่างกว้างขวาง ดังนั้น เราจึงไม่จำเป็นต้องเขียนค่าต่าง ๆ ดังรายละเอียดของตารางที่ 11.3 เราจะเขียนเฉพาะค่าสรุปคือผลรวมของแต่ละคอลัมน์ สำหรับค่า a และ b ที่ได้นั้น มาจากสูตรซึ่งได้จากสมการปกติ คือ

$$a = \bar{Y} - b\bar{X} \text{ หรือ } \frac{(\Sigma X^2)(\Sigma Y) - (\Sigma X)(\Sigma XY)}{n(\Sigma X^2) - (\Sigma X)^2} \qquad (11.3)$$

และ $b = \dfrac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$ หรือ $\dfrac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$ (11.4)

เราจะใช้สมการที่ (11.3) และ (11.4) หาค่า a และ b ดังนี้

$$a = \frac{(1261)(131) - (147)(1012)}{20(1261) - (147)^2} = \frac{16{,}427}{3611} = 4.55$$

$$b = \frac{20(1012) - (147)(131)}{20(1261) - (147)^2} = \frac{983}{3611} = 0.27$$

จะได้ค่า a และ b เท่าเดิม

เส้น Y = 4.55 + 0.27 X คือเส้นในรูป 11.6 และเส้นนี้จะให้ผลรวมของค่ากำลังสองของระยะทางแนวดิ่ง จากจุดต่าง ๆ ไปยังเส้นมีค่าน้อยที่สุด

เมื่อเราหาสมการกำลังสองน้อยที่สุดได้แล้ว เราจะพยากรณ์คะแนนผลการทำงานจาก GPA เช่น สมมุติว่าพนักงานเข้าใหม่คนหนึ่งมี GPA 10 คะแนน เราจะพยากรณ์ผลงานได้โดยการแทนค่า X = 10 ในสมการ

$$\hat{Y} = 4.55 + 0.27(10) = 7.25$$

ในเมื่อ $\hat{Y}$ คือค่า Y ที่คำนวณได้จากสมการ ซึ่งต่างจากค่า Y ที่เราสังเกตจากตัวอย่าง อันที่จริงคะแนนผลงาน 7.25 นี้ คือ **ค่าคาดหมาย**โดยกำหนดว่า GPA = 10 พึงสังเกตว่ามีหลายคนที่จบด้วย GPA 10 คะแนนนี้ และแต่ละคนน่าจะมีผลงานต่างกัน ดังนั้นอาจมองอีกด้านหนึ่งได้ว่าค่าที่พยากรณ์ได้จากสมการกำลังสองน้อยที่สุดนี้คือ **ค่าเฉลี่ย**ด้วยเหตุนี้ จึงมีหลายคนเรียกเส้นกำลังน้อยที่สุดนี้ว่า **ค่าเฉลี่ยภายใต้เงื่อนไข** หรือ conditional mean นั่นคือ ทุก ๆ จุดบนเส้นนี้ คือ ค่าเฉลี่ยของค่า Y ทั้งหลาย เมื่อกำหนดค่า X

เราเรียกสมการกำลังสองน้อยที่สุด Y = a + b X ว่า ความถดถอย Y บน X (regression of Y on X) คำว่า "ความถดถอย" หรือ regression เป็นคำที่ท่าน Francis Galton บัญญัติขึ้นในศตวรรษที่ 19 จากการศึกษาความสัมพันธ์ระหว่างความสูงของบิดา และบุตรชาย เขาพบว่าบิดาที่สูงมากมักจะมีบุตรชายสูงมากเช่นกัน ส่วนบิดาที่เตี้ยมากมักจะมีบุตรชายที่เตี้ยมากเช่นกัน อย่างไรก็ตามความสูงเฉลี่ยของบุตรชายที่มีบิดาสูงมากจะต่ำกว่าความสูงเฉลี่ยของบิดาเหล่านั้น และความสูงเฉลี่ยของบุตรชายที่มีบิดาเตี้ยมากจะสูงกว่าความสูงเฉลี่ยของบิดาเหล่านั้น Galton จึงเรียกความโน้มเอียงที่จะกลับไปสู่ความสูงเฉลี่ยของประชากรทั้งหมดอันเป็นความสูงตามปกติว่า "การถดถอย" นั่นคือเส้นถดถอย คือเส้นค่าเฉลี่ยแบบมีเงื่อนไขนั่นเอง

เส้นถดถอยที่เราได้มานี้ เป็น**เส้นถดถอยตัวอย่าง** (sample regression line) เพราะเราสร้างโดยใช้ข้อมูลจากตัวอย่างคือ พนักงาน 20 คน ถ้าเราสุ่มพนักงานมาอีกชุดหนึ่งจำนวน 20 คน แล้วหาค่า a และ b เราจะได้ค่า a และ b ที่ต่างจากที่หาไว้แล้ว ดังนั้นเราจะได้เส้นถดถอยที่ต่างไปจากเส้นเดิม ถ้าเราแทนเส้นถดถอยของประชากร ซึ่งเป็นเส้นถดถอยที่แท้จริงด้วย

$$\mu_{y/x} = \alpha + \beta X \qquad (11.5)$$

ในเมื่อ $\mu_{y/x}$ คือ ค่าที่แท้จริง (ค่าประชากร) หรือค่าเฉลี่ยของ Y เมื่อกำหนดค่า X, $\alpha$ คือจุดที่เส้น

ถดถอยประชากรตัดแกน Y และ $\beta$ คือ ความลาดชันของเส้นถดถอยของประชากร และเราจะใช้ a เป็นค่าประมาณของ $\alpha$ และ b เป็นค่าประมาณของ $\beta$ เราจึงสรุปได้ว่า เส้น $Y = a + bX$ เป็นเส้นที่ใช้ประมาณเส้น

$$\mu_{y/x} = \alpha + \beta X$$

จึงเห็นได้ว่าที่แท้จริงแล้ว เส้นถดถอยก็คือค่าเฉลี่ยภายใต้เงื่อนไขนั่นเอง และค่า Y ที่เราคำนวณได้จากสมการถดถอยตัวอย่าง ก็คือ ค่าเฉลี่ยจากตัวอย่างของ Y สำหรับค่า X ที่กำหนดให้ นั่นคือ $\mu_{y/x}$ คือ **ค่าคาดหมาย** (expected Value) หรือค่าเฉลี่ยของประชากรของ Y เมื่อกำหนดค่า X

## ข้อสังเกต

1. ตัวประมาณค่าแบบกำลังสองน้อยที่สุดคือ a และ b ของสมการถดถอยของประชากร (11.5) เป็นตัวประมาณค่าที่ไม่เอียงเฉ และมีประสิทธิภาพสูงกว่าตัวประมาณค่าซึ่งหามาโดยวิธีอื่น นอกเหนือจากวิธีกำลังสองน้อยที่สุด

2. ในการหาตัวประมาณค่าด้วยวิธีกำลังสองน้อยที่สุดต้องใช้แคลคูลัสช่วยให้

$Q = \Sigma^{n}_{=1} (Yi - (\alpha + \beta Xi))^2$

Q จะมีค่าน้อยที่สุดได้โดยมีค่า $\alpha$ และ $\beta$ ที่เหมาะสม ค่า $\alpha$ และ $\beta$ ที่จะทำให้ Q มีค่าน้อยที่สุดจะหาได้โดยวิธี partial derivatives Q with respect to $\alpha$ และ $\beta$ ดังนี้

$$\frac{\partial Q}{\partial \alpha} = \Sigma_{i=1}^{n} \frac{\partial}{\partial \alpha} (Y_i - \alpha - \beta X_i)^2 = -2\Sigma (Y_i - \alpha - \beta X_i)$$

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} (Y_\cdot - \alpha - \beta X_i)^2 = 2\sum_{i=1}^{n} X_i (Y_i - \alpha - \beta X_i)$$

เมื่อเทียบสมการทั้งคู่กับศูนย์ และใช้ค่า a และ b แทนค่า $\alpha$ และ $\beta$ ตามลำดับ จะได้

$-2\Sigma(Y_i - \alpha - \beta X_i) = 0$

$-2\Sigma X_i(y_i - \alpha - \beta X_i) = 0$

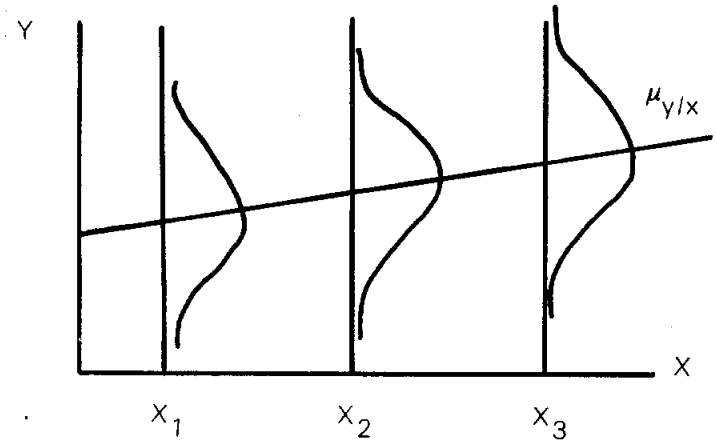
และในที่สุดจะกลายเป็นสมการปกติ (11.1) และ (11.2)

## การอนุมานเกี่ยวกับสัมประสิทธิความถดถอย $\beta$

ก่อนที่เราจะหาช่วงเชื่อมั่น และทดสอบสมมติฐานเกี่ยวกับ $\beta$ ควรทราบวิธีการหาค่าเบี่ยงเบนมาตรฐานของ Y เมื่อกำหนดค่า X

เราทราบว่าแต่ละค่าของ X จะมีค่า Y ที่สังเกตได้ซึ่งมีค่าต่าง ๆ ในลักษณะเชิงสุ่ม กล่าวคือ Y เหล่านั้นจะมีการแจกแจงความน่าจะเป็นโดยมี $\mu_{y/x}$ เป็นค่าเฉลี่ย และ $\sigma_{y/x}$ เป็นค่าเบี่ยงเบน

มาตรฐาน และยังสมมุติต่อไปอีกว่า มีการแจกแจงแบบปกติ รูปที่ 11.7 จะแสดงการแจกแจงของ Y เมื่อกำหนดค่า $X_1, X_2$ และ $X_3$ รวม 3 ประชากร มีข้อสังเกตคือ $\mu_{y/x}$ ซึ่งเป็นค่าเฉลี่ยของแต่ละประชากรคือจุดบนเส้นถดถอยประชากร




**รูปที่ 11.7** แสดงการแจกแจงความน่าจะเป็นแบบมีเงื่อนไขของ Y เมื่อกำหนดค่า X ให้ 3 ค่า

เราประมาณค่า $\mu_{y/x}$ ด้วย $\hat{Y}$ หรือค่า Y ที่คำนวณได้ , $\hat{Y}$ คือจุดบนเส้นถดถอยตัวอย่าง และเราประมาณค่า $\sigma_{y/x}$ ด้วย $S_{y/x}$ โดยที่ $S_{y/x}$ คือค่าเบี่ยงเบนมาตรฐานของ Y เมื่อกำหนด X และหาได้ดังนี้

$$S_{y/x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} \qquad (11.6)$$

ค่า n - 2 คือ degree of freedom และเหตุที่ต้องนำ 2 มาหักออกเพราะเราต้องประมาณพารามิเตอร์ 2 ตัว คือ $\alpha$ และ $\beta$

เราไม่ใคร่นิยมใช้สมการ (11.6) เพราะต้องหาค่า $\hat{Y}$ นั่นคือต้องสร้างสมการถดถอยแล้วหาผลต่างของค่า Y ที่สังเกตจากตัวอย่าง กับ Y ที่คำนวณได้ ซึ่งจะต้องหาทั้งหมด n ค่า แล้วหาผลบวกกำลังสองของค่าเบี่ยงเบน n ค่านี้อีกที สรุปแล้ว จะต้องคำนวณค่าต่าง ๆ ดังนี้

1. a = ____, b = _____
2. นำค่า a และ b ที่หาได้ สร้างสมการถดถอยตัวอย่าง

   $Y = a + bX$
3. นำค่า $X_i$ ทั้งหลายแทนค่าในสมการจะได้ $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$
4. หา $(Y_i - \hat{Y}_i)$ รวม n จำนวน

5. หา $\Sigma(Y_i - \hat{Y}_i)^2$

| X | Y | $\hat{Y}$ | $(Y - \hat{Y})$ | $(Y - \hat{Y})^2$ |
|---|---|---|---|---|
| $X_1$ | $Y_1$ | $\hat{Y}_1$ | $Y_1 - \hat{Y}_1$ | $(Y_1 - \hat{Y}_1)^2$ |
| $X_2$ | $Y_2$ | $\hat{Y}_2$ | $Y_2 - \hat{Y}_2$ | $(Y_2 - \hat{Y}_2)^2$ |
| | | | | |
| $X_n$ | $Y_n$ | $\hat{Y}_n$ | $Y_n - \hat{Y}_n$ | $(Y_n - \hat{Y}_n)^2$ |
| | | | $\Sigma = 0$ | $\Sigma(Y - \hat{Y})^2$ |

$$S_{y/x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}}$$

เรานิยมใช้สมการ (11.7) เพราะยุ่งยากน้อยกว่า (11.7)

$$S_{y/x} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n - 2}}$$

แทนค่าจากตาราง ที่ 11.2 จะได้ $= \sqrt{\frac{921 - 4.55(131) - 0.27(1012)}{20 - 2}}$

$$= \sqrt{\frac{51.72}{18}} \qquad = 1.695$$

## การทดสอบสมมติฐานและการประมาณค่าแบบช่วงสำหรับ $\alpha$ และ $\beta$

ในบทที่ 9 และ 10 เราได้ประมาณค่าและทดสอบสมมติฐานค่าพารามิเตอร์ของประชากร โดยใช้ข้อมูลจากตัวอย่างซึ่งเป็นการวิเคราะห์เมื่อมีตัวแปรเชิงสุ่มเพียงชนิดเดียว (univariate analysis) ในการประมาณค่าและทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ สัมประสิทธิ์ความถดถอย $\beta$ ก็เช่นกัน เราจะใช้สัมประสิทธิ์ความถดถอย b จากตัวอย่างเป็นตัวประมาณค่า แต่ b เป็นค่าสถิติ จึงมีค่าที่เปลี่ยนแปลงตามตัวอย่างที่สุ่มมา เช่น b อาจมีค่าต่างจาก 0 แม้ว่า $\beta$ จะมีค่าแท้จริง = 0 ถ้า $\beta = 0$ จะแสดงว่าตัวแปรคู่นั้นไม่มีความสัมพันธ์กัน การที่ b มีค่าไม่แน่นอน เราจึงจำเป็นต้อง

ทราบลักษณะการแจกแจงของ b ในตัวอย่าง ซึ่งจะพบว่า b มีการแจกแจงแบบปกติด้วยค่าเฉลี่ย- $\beta$ และความคลาดเคลื่อนมาตรฐาน $\sigma_b$ ปกติเรามักไม่ทราบค่าแท้จริงของ $\sigma_b$ จึงต้องประมาณจากตัวอย่างโดย

$$S_b = \frac{S_{y/x}}{\sqrt{\Sigma X^2 - (\Sigma X)^2/n}} \qquad (11.8)$$

และเนื่องจากเราใช้ $S_b$ แทน $\sigma_b$ และมักเป็นตัวอย่างขนาดเล็ก เราจึงต้องใช้การแจกแจงแบบ- t แทนการแจกแจงแบบปกติ

สำหรับสัมประสิทธิ์ความถดถอย $\alpha$ ซึ่งเราใช้ a เป็นค่าประมาณแบบจุด a จะมีค่าเปลี่ยนแปลงเมื่อเปลี่ยนตัวอย่างชุดใหม่ จึงควรทราบการแจกแจงของ a ด้วย ซึ่งความจริงมีว่า การแจกแจงของ a ในตัวอย่างสุ่มจะเป็นแบบปกติด้วยค่าเฉลี่ย $\alpha$ และความคลาดเคลื่อนมาตรฐาน $\sigma_a$ และปกติเรามักไม่ทราบค่าแท้จริงของ $\sigma_a$ จึงต้องประมาณจากตัวอย่างโดย

$$S_a = S_b\sqrt{\frac{\Sigma X^2}{n}} \qquad (11.9)$$

ดังนั้นจากตาราง 11.2 จะได้

$$S_b = \frac{1.695}{\sqrt{1261 - (147)^2/20}} = 0.126$$

$$S_a = 0.126\sqrt{\frac{1261}{20}} = 1.00$$

ดังนั้น ช่วงเชื่อมั่นของ $\alpha$ คือ $a \pm t(S_a)$
และ ช่วงเชื่อมั่นของ $\beta$ คือ $b \pm t(S_b)$
เปิดค่า t จากตารางที่ $\nu = (n - 2)$ และ $\alpha/2$
ดังนั้น 95% ช่วงเชื่อมั่นของ $\alpha$ คือ

$$a \pm t_{18,.025}(S_a) = 4.55 \pm (2.101)(1.0) = 2.449,\ 6.651$$

และ 95% ช่วงเชื่อมั่นของ $\beta$ คือ

$$b \pm t_{18,.025}(S_b) = 0.27 \pm (2.101)(.126) = 0.0053,\ 0.5347$$

ถ้า n > 30 ให้ใช้ค่า Z แทนค่า t

ส่วนวิธีการทดสอบสมมุติฐานเกี่ยวกับ $\alpha$ และ $\beta$ ก็ใช้วิธีการในบทที่ 10 นั่นคือ ใช้การทดสอบแบบ t เมื่อ n ≤ 30 ค่าสถิติที่คำนวณได้จะมีการแจกแจงแบบ t ด้วย df = n − 2 และเมื่อ n > 30 ใช้ค่า Z แทน t นั่นคือตัวสถิติที่ใช้ทดสอบ $\alpha$ คือ

$$T = \frac{a - \alpha}{S_a} \qquad (11.10)$$

และตัวสถิติที่ใช้ทดสอบ $\beta$ คือ

$$T = \frac{b - \beta}{S_b} \qquad (11.11)$$

## การทดสอบความสัมพันธ์เชิงเส้นระหว่าง X และ Y

จุดประสงค์เบื้องต้นของการศึกษาความถดถอยคือ การพยากรณ์ค่า Y โดยกำหนดค่า X แต่เราควรสำรวจให้แน่ใจก่อนว่า ตัวแปรคู่นั้นมีความสัมพันธ์กันหรือไม่ ถ้าตัวแปรคู่นั้นมีความสัมพันธ์ที่แท้จริงเป็นแบบเชิงเส้น $\beta$ จะมีค่าที่ต่างจากศูนย์ ดังนั้น ในการทดสอบ เราจึงตั้งสมมติฐานว่างเปล่าว่า X และ Y ไม่มี ความสัมพันธ์เชิงเส้น นั่นคือ Ho : $\beta = 0$ ส่วนสมมติฐานรอง Ha เราจะแย้งได้หลายอย่าง ดังนี้

Ha : $\beta \neq 0$ หมายความว่า X และ Y มีความสัมพันธ์เชิงเส้น

หรือ

Ha : $\beta > 0$ หมายความว่า X และ Y มีความสัมพันธ์เชิงเส้นและความสัมพันธ์นั้นมีทิศทางเดียวกัน หรือเป็นแบบเชิงบวก

หรือ Ha : $\beta < 0$ หมายความว่า X และ Y มีความสัมพันธ์เชิงเส้น และความสัมพันธ์นั้นมีทิศทางตรงข้ามกัน หรือเป็นแบบนิเสธ

---

**ตัวอย่าง** จากข้อมูลในตาราง 11.2 จงทดสอบว่า คะแนนการทำงาน และ GPA เมื่อจบการศึกษามีความสัมพันธ์เชิงเส้นหรือไม่? $\alpha = .05$

1. Ho : $\beta = 0$ (คะแนนการทำงาน และ GPA ไม่มีความสัมพันธ์เชิงเส้น)
2. Ha : $\beta \neq 0$ (คะแนนการทำงาน และ GPA มีความสัมพันธ์เชิงเส้น)
3. $\alpha = .05$
4. จะปฏิเสธ Ho เมื่อ $T > t_{18,.025} = 2.101$ หรือ $T < -2.101$
5. $T = \frac{b - \beta}{S_b} = \frac{0.27 - 0}{0.126} = 2.143$

6. $T = 2.143 > 2.101$ จึงปฏิเสธ Ho และยอมรับ Ha จึงสรุปว่า มีหลักฐานพอเพียงที่ชี้แนะว่า GPA เมื่อจบการศึกษา และคะแนนผลงานมีความสัมพันธ์เชิงเส้น

**หมายเหตุ** ถ้าตรวจดู 95% ช่วงเชื่อมั่นของ $\beta$ คือ

$$0.0053 < \beta < 0.5347$$

จะเห็นว่า $\beta = 0$ ไม่อยู่ในช่วงเชื่อมั่น โดยเราใช้ค่า $\alpha = .05$ เท่ากัน แสดงว่าต้องปฏิเสธ Ho

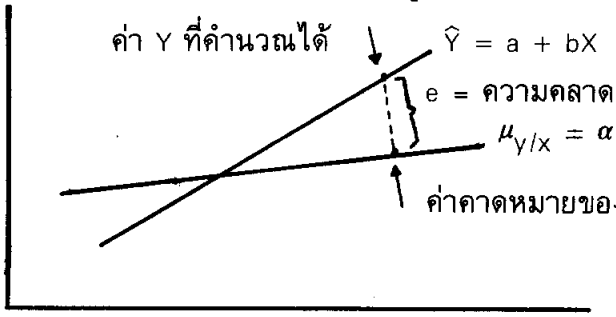
## การประมาณค่า $\mu_{y/x}$

การประมาณค่าเฉลี่ยแบบมีเงื่อนไข $\mu_{y/x}$ เป็นเทคนิคสำคัญที่นิยมใช้ในธุรกิจและการจัดการ เช่น ถ้ารายได้จากการขายสินค้า Y มีความสัมพันธ์กับค่าใช้จ่ายโฆษณา X ดังนั้น เราจะสามารถประมาณจำนวนขายโดยเฉลี่ยสำหรับงบค่าโฆษณาที่กำหนดให้ หรือถ้าผลงานของพนักงาน Y มีความสัมพันธ์กับคะแนนความถนัด X เราก็จะสามารถประมาณคะแนนผลงานโดยเฉลี่ยจากคะแนนความถนัดที่กำหนดให้ หรือถ้าทราบว่า GPA ระดับมัธยมปลาย X และ GPA ระดับมหาวิทยาลัย Y เราก็จะสามารถประมาณ GPA โดยเฉลี่ยของระดับมหาวิทยาลัยจากคะแนน GPA ระดับมัธยมปลายที่กำหนดให้ ดังนั้น เราจึงควรทราบวิธีการสร้างช่วงเชื่อมั่นของค่าเฉลี่ยที่ประมาณได้ ณ ค่า X ที่กำหนดให้

ตัวประมาณค่าที่ไม่เอียงเฉของ $\mu_{y/x}$ คือ $\hat{Y}$ และการแจกแจงตัวอย่างของ $\hat{Y}$ จะเป็นแบบโค้งปกติด้วยค่าเฉลี่ย $\mu_{y/x}$ และค่าเบี่ยงเบนมาตรฐาน $\sigma_{\hat{y}}$ แต่เรามักไม่ทราบค่าแท้จริงของ $\sigma_{\hat{y}}$ จึงใช้ $S_{\hat{y}}$ ประมาณ

$$S_{\hat{y}} = S_{y/x}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n}} \qquad (11.12)$$

โปรดสังเกตว่า $S_{\hat{y}}$ คือค่าเบี่ยงเบนมาตรฐานของเส้นถดถอยตัวอย่างที่เบี่ยงเบนไปจากค่าจริง คือ เส้นถดถอยประชากร ส่วน $S_{Y/x}$ คือค่าเบี่ยงเบนมาตรฐานของค่าสังเกต Y ที่เบี่ยงเบนไปจากเส้นถดถอยตัวอย่าง $\hat{Y}$ ดังแสดงในรูปที่ 11.7 ความแตกต่างของ $\hat{Y}$ กับ $\mu_{y/x}$ คือ sampling error




**รูปที่ 11.7** แสดงค่า Y ที่คำนวณได้จากสมการถดถอยตัวอย่าง และค่าคาดหมายของ Y

ดังนั้น $(1 - \alpha)\,100\%$ ช่วงเชื่อมั่นของ $\mu_{y/x}$ คือ

$$\hat{Y} \pm t_{\alpha/2,\,\nu} S_{\hat{y}} \qquad (11.13)$$

$\nu = n - 2$

ถ้า $n > 30$ เปิดตาราง Z แทนตาราง t

ตัวอย่าง จากข้อมูลในตารางที่ 11.2 จงหา 95% ช่วงเชื่อมั่นของค่าคาดหมายของ $\hat{Y}$ หรือ $\mu_{y/x}$ เมื่อ

(1) $X = 4$

(2) $X = 7.35$

(3) $X = 10$

$S_{y/x} = 1.695,\ t_{18,.025} = 2.101,\ \bar{X} = \frac{147}{20} = 7.35$

| $X$ | $\hat{Y} = 4.55 + 0.27X$ | $t_{(18,.025)}$ | $S_{y/x}$ | $\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n}}$ |
|---|---|---|---|---|
| 4 | 5.63 | 2.101 | 1.695 | $\sqrt{\frac{1}{20} + \frac{(4 - 7.35)^2}{180.55}} = .3349$ |
| 7.35 | 6.5345 | 2.101 | 1.695 | $\sqrt{\frac{1}{20} + \frac{(7.35 - 7.35)^2}{180.55}} = .2236$ |
| 10 | 7.25 | 2.101 | 1.695 | $\sqrt{\frac{1}{20} + \frac{(10 - 7.35)^2}{180.55}} = .2982$ |

ดังนั้น 95% ช่วงเชื่อมั่นของ $\mu_{y/x}$ มีดังนี้

1) เมื่อ $X = 4$, $5.63 \pm (2.101)(1.695)(.3349) = 4.437,\ 6.823$

หรือ $4.437 < \mu_{y/x\,=\,4} < 6.823$

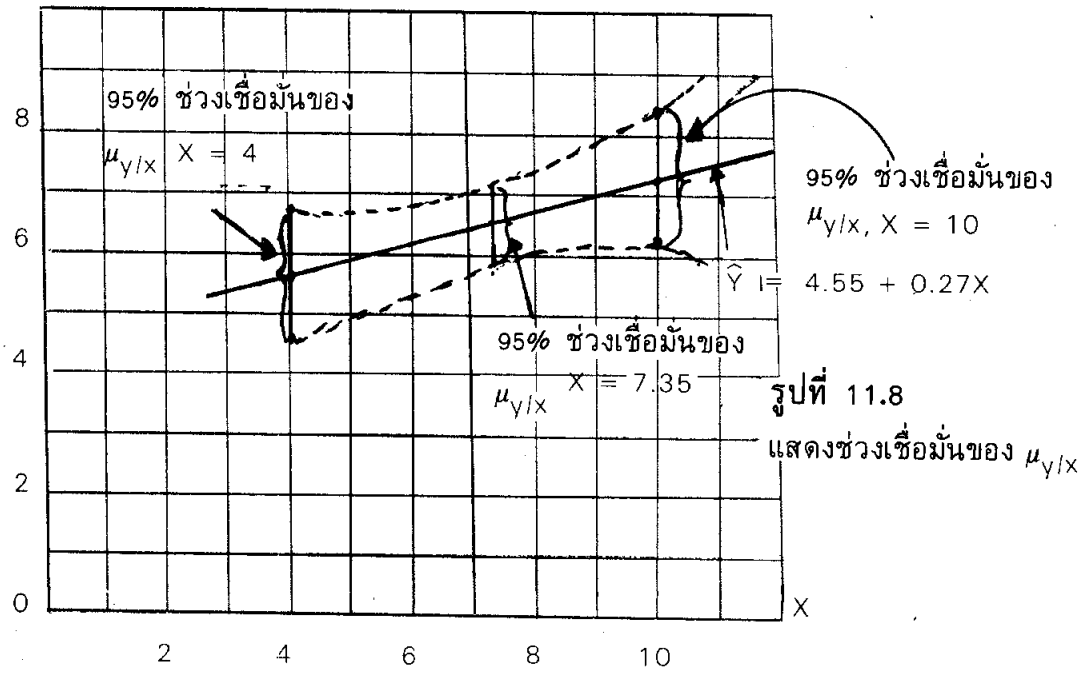
2) เมื่อ $X = 7.35$, $6.5345 \pm (2.01)(1.695)(.2236) = 5.738,\ 7.331$

หรือ $5.738 < \mu_{y/x\,=\,7.35} < 6.823$

3) เมื่อ $X = 10$, $7.25 \pm (2.101)(1.695)(.2982) = 6.188,\ 8.312$

หรือ $6.188 < \mu_{y/x\,=\,10} < 8.312$

ถ้าสังเกตให้ดีจะเห็นว่า ช่วงเชื่อมั่นของ $\mu_{y/x}$ จะกว้างออกเมื่อ X ห่างจาก $\bar{X}$ ซึ่งชี้ให้เห็นว่าค่าเฉลี่ยที่ประมาณได้มีความเชื่อถือได้น้อยลง จากตัวอย่างระหว่างค่า X ต่าง ๆ 3 ค่า $S_{\hat{y}}$ จะมีค่าต่ำสุดเมื่อ $X = \bar{X}$ และจะมีค่าสูงขึ้นเรื่อยเมื่อ X ออกห่าง $\bar{X}$ ดังรูป 11.8




**รูปที่ 11.8**
แสดงช่วงเชื่อมั่นของ $\mu_{y/x}$

## การประมาณค่าที่แท้จริงของ Y

(Estimation of the actual Y value)

$\mu_{y/x}$ คือ ค่าเฉลี่ยแบบมีเงื่อนไข หรือค่าคาดหมายของ Y เมื่อกำหนดค่า X แต่มีบ่อยครั้งที่เราต้องการพยากรณ์บางค่าของ Y เช่น ชาวนาผู้หนึ่งต้องการประมาณผลผลิตข้าว Y ในปีหนึ่งเมื่อใส่ปุ๋ย X จำนวนหนึ่ง, ประธานบริษัทอยากทราบจำนวนขาย Y ในไตรมาสต่อไปจากงบค่าโฆษณา-X ที่กำหนดให้ ให้ $Y_a$ หรือ $Y_{actual}$ คือ ค่าพารามิเตอร์ Y ที่ต้องการประมาณเมื่อกำหนดค่า X ดังนั้น ค่าประมาณแบบจุดของ $Y_a$ ก็คือ $\hat{Y}$ นั้นเอง แต่ความคลาดเคลื่อนมาตรฐานในการประมาณค่า $Y_a$ จะสูงกว่า $\mu_{y/x}$ โดยที่

$$\sigma_{y_a} = \sigma_{y/x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n}}$$

และถ้าไม่ทราบค่า $\sigma_{y/x}$ ต้องประมาณด้วย $S_{y/x}$ จะได้

$$S_{Y_a} = S_{y/x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - (\Sigma X)^2/n}} \qquad (11.14)$$

ดังนั้น $(1 - \alpha)$ 100% ช่วงเชื่อมั่นของ $Y_a$ คือ

$$\hat{Y} \pm t_{\alpha/2,\,(n-2)} S_{Y_a} \qquad (11.15)$$

---

## ตัวอย่าง

จากตาราง 11.2 - 11.3 ถ้าพนักงานผู้หนึ่งมี GPA เมื่อจบการศึกษาเป็น 10 จงหา 95%ช่วงเชื่อมั่นของคะแนนผลงานที่แท้จริงของพนักงานผู้นั้นในการทดสอบครั้งต่อไป

เมื่อ $X = 10$, $\hat{Y} = 4.55 + 0.27(10) = 7.25$ , $t_{18,.025} = 2.101$

$$S_{Y_a} = 1.695\sqrt{1 + \frac{1}{20} + \frac{(10 - 7.25)^2}{180.55}} = 1.7687$$

ดังนั้น 95% ช่วงเชื่อมั่นของ Ya คือ

$$7.25 \pm 2.101(1.7687) = 3.534,\ 10.966$$

หรือ

$$3.534 < Y_a < 10.966$$

พึงสังเกตว่าช่วงเชื่อมั่นของ $Y_a$ จะกว้างกว่า $\mu_{y/x}$ เมื่อกำหนดค่า X เท่ากันคือ $X = 10$

---

## ข้อสมมุติที่สำคัญในการศึกษาความถดถอย

ในการวิเคราะห์ความถดถอย จะต้องมีข้อสมมุติ (assumptions) 5 ประการ ดังนี้

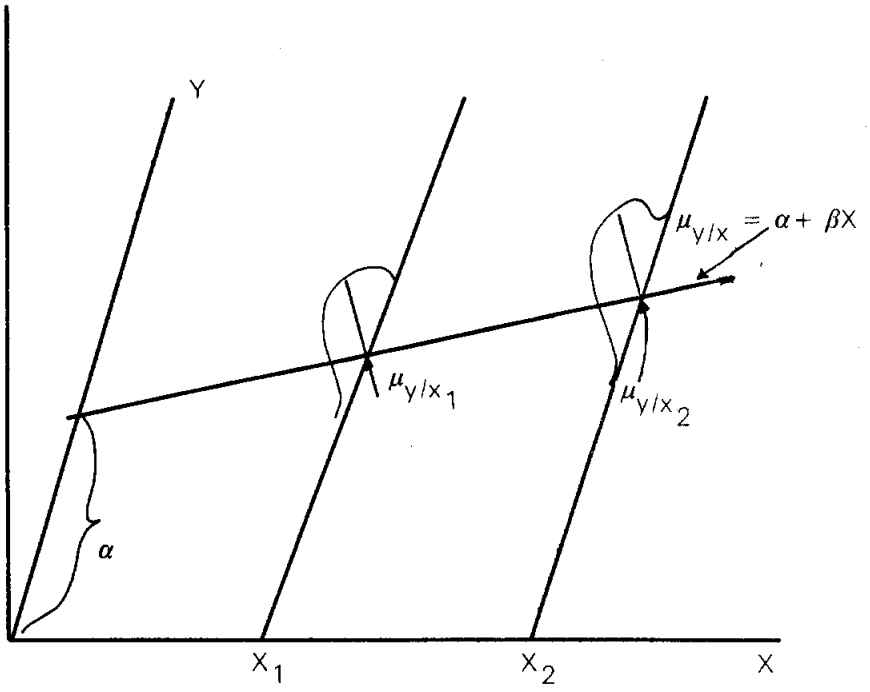
**1. ความสัมพันธ์เชิงเส้น** (linearity)

คือ สมมุติว่ามีความสัมพันธ์เชิงเส้นระหว่างค่าเฉลี่ยของประชากร คือ $\mu_{y/x}$ กับค่าต่าง ๆ ของ X นั่นคือ ค่าเฉลี่ยของประชากร Y ทั้งหลายจะอยู่บนเส้นตรง นั่นคือ

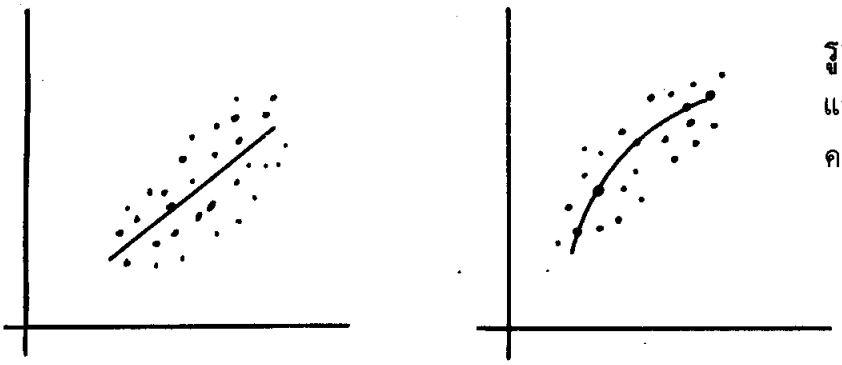
$$\mu_{y/x} = \alpha + \beta X$$

ในเมื่อ $\alpha$ และ $\beta$ คือค่าพารามิเตอร์ และมี a และ b เป็นตัวประมาณค่าที่มีประสิทธิภาพ และไม่เอียงเฉ

รูปที่ 11.9 และ 11.10 จะแสดงคุณสมบัติของข้อสมมุตินี้ เมื่อมีประชากรของ Y จำนวน 2 ประชากร ส่วนรูปที่ 11.10 เป็น scatter diagram 2 อัน อันซ้ายมือแสดงว่า ข้อสมมุติว่ามีความสัมพันธ์เชิงเส้นน่าจะเป็นจริง ส่วนรูปซ้ายมือแสดงความสัมพันธ์แบบเส้นโค้ง




**รูปที่ 11.9**
แสดงข้อสมมุติ
ความสัมพันธ์เชิงเส้นตรง



แสดงความสัมพันธ์
เชิงเส้นตรง

แสดงความสัมพันธ์
แบบเส้นโค้ง

**รูปที่ 11.10**
แสดงข้อสมมุติ
ความสัมพันธ์เชิงเส้นตรง

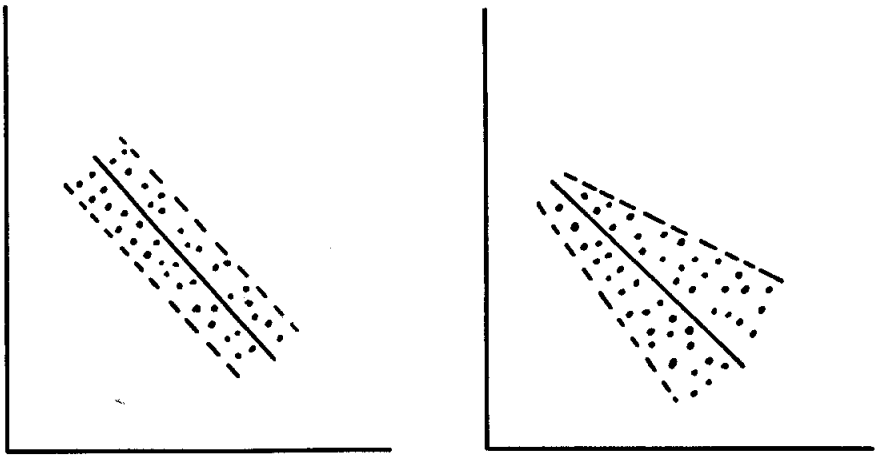
**2. ความแปรปรวนเป็นเอกภาพ** (equal variance)

แต่ละค่าของ X จะมีประชากร Y ซึ่งเรียกว่า subpopulation Y/X นั่นคือจะมีค่า Y หลายค่า แต่ประชากรย่อยเหล่านี้จะมีความแปรปรวนเท่ากัน นั่นคือ $\sigma^2_{y/x_1} = \sigma^2_{y/x_2} = \ldots = \alpha^2_{y/x_n}$

$= \sigma^2_{y/x}$

บางครั้งจะเรียกคุณสมบัติข้อนี้ว่า homocedasticity ในรูปที่ 11.11 แสดง scatter diagram 2 อัน รูปซ้ายมือแสดงว่าข้อสมมุติของความเป็นเอกภาพน่าจะเป็นจริง ส่วนรูปขวามือเห็นชัดว่าเมื่อ X มีค่าสูงขึ้น ความแปรปรวนของ Y จะมากขึ้นด้วย

**รูปที่** 11.11
แสดงข้อสมมุติ
ความแปรปรวน
เป็นเอกภาพ



ความแปรปรวน
เป็นเอกภาพ

ความแปรปรวน
ไม่เท่ากันทั้งหมด

**3. ความเป็นอิสระกัน** (independence)

หมายถึงค่าของ Y ในแต่ละค่าของ X จะเป็นอิสระกัน น่าเสียดายที่ไม่สามารถแสดงข้อสมมุตินี้ด้วย scatter diagram

**4. กำหนดค่า X ได้**

สามารถกำหนดค่าต่าง ๆ ของตัวแปรอิสระ คือ $X_1, X_2, \ldots, X_n$ ได้ล่วงหน้าดังนั้น X จึงเป็น fixed variable และ Y เป็น random variable

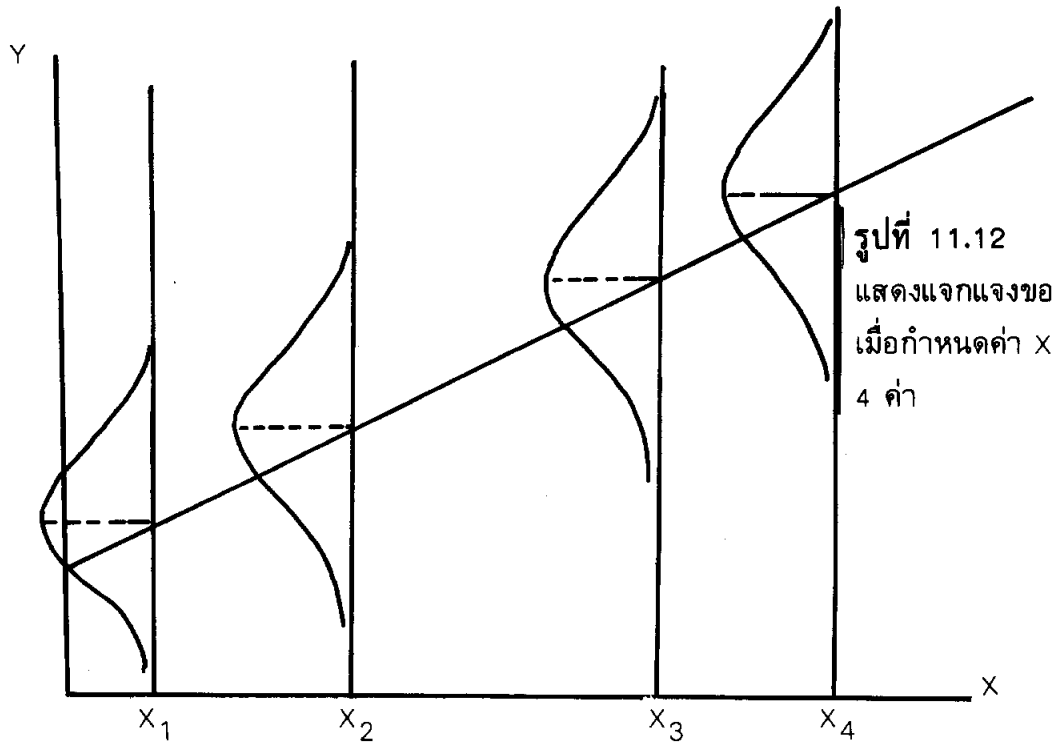
**5. การแจกแจงแบบปกติ** (normality)

ข้อสมมุตินี้ กำหนดให้รูปร่างของการแจกแจงของ Y ในแต่ละค่าของ X เป็นแบบโค้งปกติ

เมื่อนำคุณสมบัติทั้ง 5 ประการ รวมกันจะได้ตัวแบบดังนี้

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

ในเมื่อ $\epsilon_i$ คือ ความคลาดเคลื่อน (error term) และ $\epsilon_i$ จะเป็นตัวแปรเชิงสุ่มที่มีการแจกแจงแบบปกติ ซึ่งมีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนคือ $\sigma^2_{y/x}$ และแสดงได้โดยรูป 11.12




**รูปที่ 11.12** แสดงแจกแจงของ Y เมื่อกำหนดค่า X ให้ 4 ค่า

## แบบฝึกหัด

11.6 จงใช้ข้อมูลที่กำหนดให้

ก) พล็อตใน scatter diagram

ข) สร้างสมการประมาณค่าที่ใช้แทนข้อมูลได้ดีที่สุด

ค) พยากรณ์ค่า Y เมื่อ X = 4, 9 และ 12

| X | 7 | 10 | 8 | 5 | 11 | 3 | 7 | 11 | 12 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 2.0 | 3.0 | 2.4 | 1.8 | 3.2 | 1.5 | 2.1 | 3.8 | 4.0 | 2.2 |

11.7 จงใช้ข้อมูลที่กำหนดให้ข้างล่าง

ก) สร้าง scatter diagram

ข) สร้างสมการถดถอย

ค) พยากรณ์ค่า Y เมื่อ X = 12, 14 และ 18

| X | 20 | 11 | 15 | 10 | 17 | 19 |
|---|---|---|---|---|---|---|
| Y | 5 | 15 | 14 | 17 | 8 | 9 |

11.8

| X | 56 | 48 | 42 | 58 | 40 | 39 | 50 |
|---|---|---|---|---|---|---|---|
| Y | 9.5 | 7.5 | 7.0 | 9.5 | 6.2 | 6.6 | 8.7 |

ก) หาเส้นตรงที่ปรับเข้ากับข้อมูลได้ดีที่สุด

ข) หาความคลาดเคลื่อนมาตรฐานของค่าประมาณ (standard error of estimate หรือ $S_{y/x}$)

ค) หา 95% ช่วงเชื่อมั่นของค่าพยากรณ์ เมื่อ X = 44

11.9 อุดมเป็นผู้จัดการร้านขายเครื่องไฟฟ้า เขาต้องการทราบว่า เงินที่เขาใช้โฆษณาโดยการซื้อเวลาจากโทรทัศน์และวิทยุมีความสัมพันธ์กับจำนวนขายอย่างไร เขาได้เก็บข้อมูลคือ เวลาโฆษณาเป็นนาที (X) และจำนวนสินค้าหลักที่ขายได้ต่อสัปดาห์ (Y) ของ 7 สัปดาห์ก่อน ดังนี้

| X (นาที) | 25 | 18 | 32 | 21 | 35 | 28 | 30 |
|---|---|---|---|---|---|---|---|
| Y | 16 | 11 | 20 | 15 | 26 | 32 | 20 |

ก) จงสร้างสมการถดถอย

ข) จงหา standard error of estimate

ค) หาช่วงเชื่อมั่นของค่าพยากรณ์ของ Y เมื่อ X = 27, $\alpha$ = .10

11.10 ที่ปรึกษาฝ่ายผลิตของโรงงานหนึ่งให้คำแนะนำว่า โรงงานควรพิจารณาแจกงานซึ่งต้องใช้แรงงานสูงให้เหมาะสมกับอายุของพนักงาน เขาได้สุ่มพนักงานมา 10 คน และได้บันทึกระยะเวลาที่สามารถแบกหามของหนัก ดังนี้

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| อายุ | 42 | 27 | 36 | 25 | 22 | 39 | 57 | 19 | 33 | 30 |
| นาฑีของงานหนัก | 2 | 7 | 5 | 9 | 10 | 4 | 4 | 8 | 6 | 5 |

ก) จงพล็อตข้อมูล

ข) สร้างสมการแสดงความสัมพันธ์ของอายุ และความสามารถของร่างกาย

ค) เราจะคาดว่า พนักงานที่มีอายุ 30 ปี คนหนึ่งจะสามารถแบกหามของหนักได้นานเท่าใด

11.11 จงแสดงว่า เส้น $Y = a + bX$ และ $Y = \bar{Y} + b(X - \bar{X})$ คือเส้นเดียวกัน

11.12 บริษัทประกันชีวิตแห่งหนึ่งต้องการทราบความสัมพันธ์ของประสพการณ์กับจำนวนขาย จึงสุ่มพนักงานมา 9 คน ให้ X คือจำนวนปีที่ทำงานและ Y คือจำนวนขายต่อปีมีหน่วยเป็น 100,000 บาท ได้ข้อมูลดังนี้

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| X | 2 | 1 | 3 | 3 | 4 | 5 | 6 | 5 | 7 |

ก. จงประมาณจำนวนขายต่อปีของพนักงานผู้หนึ่งซึ่งทำงานมา 10 ปี

ข. มีหลักฐานพอเพียงที่แสดงว่า $\beta$ ไม่เป็น 0 ที่ $\alpha = .05$ ไหม?

ค. จงหา 95% ช่วงเชื่อมั่นของ $\mu_{y/x}$ เมื่อ $X = 4.5$ และ $X = 10$

ง. จงสร้าง 95% ช่วงเชื่อมั่นของ $Y_a$ เมื่อ $X = 10$

11.13 ในการวิเคราะห์ความสัมพันธ์ระหว่างคะแนนผลงานของวิศวกร 10 คน และระยะเวลาที่ใช้วิชาชีพ ได้ข้อมูลดังนี้

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y = คะแนนผลงาน | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 9 |
| X = อายุการใช้วิชาชีพ | 1 | 3 | 2 | 5 | 5 | 4 | 7 | 6 | 9 | 8 |

ก. จงหาเส้นกำลังสองน้อยที่สุดเพื่อ "fit" ข้อมูล

ข. จงประมาณคะแนนผลงานสำหรับผู้ใช้วิชาชีพ 10 ปี

ค. ทดสอบความสัมพันธ์เชิงเส้นระหว่างคะแนนผลงาน และอายุการใช้วิชาชีพโดยใช้ $\alpha = 0.01$

ง. จงหา 99% ช่วงเชื่อมั่นของ $\mu_{y/x}$ เมื่อ X = 5 และ X = 1

จ. จงหา 99% ช่วงเชื่อมั่นของ $Y_a$ เมื่อ X = 10

11.14 สถิติค่าใช้จ่ายซ่อมแซมเครื่องจักรชนิดเดียวกันแต่มีอายุการใช้งานต่างกันจำนวน 6 เครื่อง โดยให้ X คือมีอายุการใช้งาน และ Y คือค่าบำรุงรักษามีหน่วยเป็น 1,000 บาท มีดังนี้

| เครื่องจักร | X | Y |
|---|---|---|
| 1 | 2 | 70 |
| 2 | 1 | 40 |
| 3 | 3 | 100 |
| 4 | 2 | 80 |
| 5 | 1 | 30 |
| 6 | 3 | 100 |

ก. จงหาสมการถดถอย Y บน X

ข. จะต้องใช้ค่าบำรุงรักษาเท่าใดสำหรับเครื่องจักรที่มีอายุ 4 ปี

ค. จงทดสอบความสัมพันธ์เชิงเส้นระหว่างอายุการใช้งาน และค่าบำรุงรักษา, $\alpha$ = .01

ง. จงหาช่วงเชื่อมั่น 99% ของ $\mu_{y/x}$ เมื่อ X = 2

จ. จงหาช่วงเชื่อมั่น 99% ของ $Y_a$ เมื่อ X = 2

11.15 สถิติการลาป่วยต่อปีของพนักงาน 5 คน มีดังนี้

| จำนวนปีทำงาน (X) | จำนวนวันลาป่วยต่อปี (Y) |
|---|---|
| 15 | 10 |
| 9 | 16 |
| 13 | 14 |
| 11 | 15 |
| 12 | 15 |

ก. จงสร้างเส้นถดถอย

ข. จงหาจำนวนวันลาป่วยต่อปีของพนักงานที่ทำงานมา 14 ปี

ค. มีเหตุผลเพียงพอที่จะรับว่า X และ Y มีความสัมพันธ์เชิงเส้นไหม? $\alpha = .05$

ง. จงหา 95% ช่วงเชื่อมั่นของ $\mu_{Y/x}$ เมื่อ X = 12 และ X = 15

จ. จงหา 95% ช่วงเชื่อมั่นของ $Y_a$ เมื่อ X = 15

11.16 ให้ X คือคะแนนทดสอบความถนัด Y คือจำนวนผลผลิตต่อชั่วโมงข้อมูลสรุปจากพนักงาน 10 คน มีดังนี้

$n = 10 \quad \Sigma X = 550 \quad \Sigma Y = 680$

$\Sigma XY = 45{,}900 \quad \Sigma X^2 = 38{,}500 \quad \Sigma Y^2 = 56{,}000$

ก) จงหาสมการถดถอย Y บน X

ข) มีหลักฐานพอเพียงที่จะสรุปว่า สัมประสิทธิ์ความถดถอย $\beta$ มีค่าแตกต่างจาก 0 ด้วย $\alpha = 0.02$ ไหม?

ค) จงหา 98% ช่วงเชื่อมั่นของ $\mu_{Y/x}$ เมื่อ (1) X = 40, (1) X = 55, (3) X = 70

ง) จงหา 98% ช่วงเชื่อมั่นของ $Y_a$ เมื่อ (1) X = 40, (2) X = 55, (3) X = 70

จ) จงอธิบายความแตกต่างของคำตอบในข้อ (ค) และ (ง)

---

## ข้อควรระวังในการสรุปผลการทดสอบ

ในการทดสอบสมมติฐานเกี่ยวกับ $\alpha$ และ $\beta$ โดยสมมุติว่า เรามีข้อสมมุติของความถดถอยที่สมบูรณ์ ได้แก่ ข้อสมมุติเกี่ยวกับการแจกแจงแบบปกติ, ความเป็นอิสระ, ความแปรปรวนเป็นเอกภาพ เป็นต้น ควรเข้าใจความหมายของข้อสรุป ดังนี้

## การทดสอบว่าเส้นถดถอยไม่มีความชัน

การทดสอบที่สำคัญที่สุดในการศึกษาความถดถอยคือ การทดสอบว่า ความลาดชันของเส้นถดถอยประชากรมีค่าต่างจาก 0 หรือไม่ ซึ่งเทียบเท่ากับการทดสอบว่า X ช่วยพยากรณ์ - Y ได้ดีโดยการใช้ตัวแบบเชิงเส้นหรือไม่ นั่นคือการทดสอบ Ho : $\beta = 0$ ในการสรุปผลการทดสอบ นอกจากจะมีโอกาสเกิดความผิดประเภทที่ 1 (คือการปฏิเสธ Ho ซึ่งเป็นจริง) และความผิดประเภทที่ 2 (คือการยอมรับ Ho ซึ่งเป็นเท็จ) เราจะอธิบายความหมายเมื่อเรายอมรับหรือปฏิเสธ Ho ได้ดังนี้

1. **เมื่อยอมรับ** Ho : $\beta = 0$ (หรือไม่ปฏิเสธ Ho) อาจมีความหมายได้ 2 อย่าง ดังนี้

(ก) X และ Y อาจมีความสัมพันธ์ที่แท้จริงเป็นแบบเชิงเส้นตรง คือตัวแบบ $\mu_{Y/X} = \alpha + \beta X$ ใช้ได้ แต่X ช่วยพยากรณ์ Y ได้น้อยมาก หรืออาจไม่ช่วยเลย นั่นคือ จะใช้ $\bar{Y}$ หรือ $\bar{Y} + \hat{\beta}\ (X - \bar{X})$ พยากรณ์ค่า Y จะได้ผลเท่ากัน ดังรูป 11.13 (ก) หรือ

(ข) ความสัมพันธ์ระหว่าง X และ Y ไม่ใช่เส้นตรง นั่นคือตัวแบบที่แท้จริงอาจเป็นรูปสมการกำลังสอง หรือกำลังสาม หรือสมการที่ซับซ้อนกว่าสมการเส้นตรง ดังรูปที่ 11.13 (ข)

ดังนั้น เมื่อเอาข้อ (ก) และ (ข) รวมกัน จึงสรุปได้ว่า เมื่อยอมรับ Ho : $\beta = 0$ (อย่าลืมว่าการยอมรับ Ho ใด ๆ หมายความเพียง" ยังมีหลักฐานไม่เพียงพอที่จะปฏิเสธ Ho) จะแสดงว่าฟังก์ชั่นเส้นตรงไม่ใช่ตัวแบบที่ดีที่สุด และไม่ช่วยมากนักในการพยากรณ์ Y

2. **เมื่อปฏิเสธ** Ho : $\beta = 0$ มีความหมายดังนี้

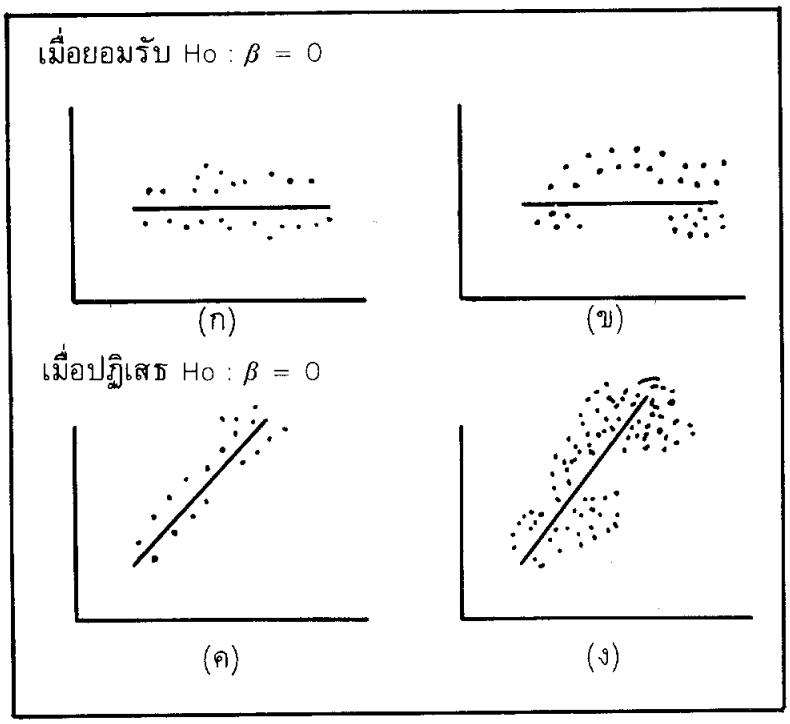
(ก) X ให้ข่าวสารที่มีนัยสำคัญสำหรับพยากรณ์ Y นั่นคือตัวแบบ $\bar{Y} + \hat{B}\ (X - \bar{X})$ ดีกว่า $\bar{Y}$ เฉย ๆ ในการพยากรณ์ค่า Y ดังรูปที่ 11.13 (ค)

หรือ

(ข) อาจจะมีตัวแบบที่ดีกว่า เช่น สมการเส้นโค้ง ดังรูปที่ 11.13 (ง) นั่นคือต้องมีค่า X กำลังหนึ่งอยู่ในตัวแบบด้วย

ดังนั้น เมื่อรวมข้อ (ก) และ (ข) ด้วยกัน จะกล่าวได้ว่า เมื่อเราปฏิเสธ Ho : $\beta = 0$ หมายความว่า สมการเส้นตรงปรับเข้ากับข้อมูลได้ดีกว่าสมการที่ไม่มี X อยู่เลย แต่ในขณะเดียวกันสมการเส้นตรงนั้น อาจเป็นเพียง "เส้นตรงโดยประมาณ" แต่ความสัมพันธ์ที่แท้จริงอาจเป็นแบบ "ไม่ใช่เส้นตรง หรือ Nonlinear" ก็ได้

สรุปได้ว่า ไม่ว่าเราจะยอมรับหรือปฏิเสธ Ho ก็ตาม อาจมีกรณีที่สมการเส้นตรงยังไม่เหมาะสม และยังมีเส้นโค้งอื่นที่อธิบายความสัมพันธ์ระหว่าง X และ Y ได้ดีกว่า




**รูปที่** 11.13
อธิบายการทดสอบ
Ho : $\beta$ = 0

## การทดสอบว่าจุดตัดที่แกน Y เป็นศูนย์

คือการทดสอบว่าเส้นถดถอยประชากรผ่านจุดกำเนิดหรือไม่ หรือ Ho : $\alpha$ = 0 ถ้ายังปฏิเสธ Ho ไม่ได้ เราอาจเอาค่า $\alpha$ ออกจากตัวแบบได้ ถ้าเรามีหลักฐานเชื่อได้ว่า เส้นตรงผ่านจุดกำเนิด และควรมีการเก็บข้อมูลสำหรับค่า X ที่อยู่ใกล้จุดกำเนิดด้วย แต่ปกติเราไม่ค่อยสนใจทดสอบสมมติฐานนี้ และเราไม่ค่อยเก็บข้อมูลที่อยู่ใกล้จุดกำเนิด เช่น ความดันโลหิต (Y) และอายุ (X) เราย่อมไม่สนใจความดันโลหิต เมื่อ X = 0 และเราไม่ค่อยเก็บข้อมูลสำหรับ X ที่มีค่าใกล้ 0

## การวิเคราะห์ความถดถอยด้วย ตารางวิเคราะห์ความแปรปรวน

เราทราบวิธีสร้างตารางวิเคราะห์ความแปรปรวนสำหรับทดสอบค่าเฉลี่ยของประชากรปกติหลาย ๆ ประชากรว่าต่างกันหรือไม่ โดยการแบ่งความผันแปรทั้งหมด หรือ SST ออกเป็นส่วน ๆ ตามที่มาของความผันแปร ในการศึกษาความถดถอยก็เช่นกัน SST จะมาจากความผันแปร 2 ส่วน คือ ความผันแปรที่อธิบายได้ว่าเนื่องจากตัวแปรอิสระ X หรือเนื่องจากความถดถอย และใช้ SSR และความผันแปรที่เหลือ หรือที่อธิบายไม่ได้ หรือความคลาดเคลื่อน เรียกว่า SSE

ดังนั้น SST = SSR + SSE

$$SST = \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2 \quad \text{มี } n - 1 \text{ df}$$

$$SSR = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad \text{มี } 1 \text{ df}$$

$$SSE = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad \text{มี } n - 2 \text{ df}$$

**เมื่อกระจายสูตรนิยามเหล่านี้ จะได้สูตรที่ใช้เครื่องคำนวณ ดังนี้**

$$SST = \Sigma Y^2 - (\Sigma Y)^2/n \quad \text{หรือ } \Sigma Y^2 - n(\bar{Y})^2$$

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

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

**หรือ**

$$SSR = b\Sigma(X - \bar{X})(Y - \bar{Y})$$

**หรือ**

$$SSR = b^2\Sigma(X - \bar{X})^2 = b^2(\Sigma X^2 - (\Sigma X)^2/n)$$

**หรือ**

$$SSR = SST - SSE$$

**ส่วน**

$$SSE = SST - SSR$$

**ตารางที่** 11.4 แสดงตารางวิเคราะห์ความแปรปรวนของความถดถอย

| ที่มาของความผันแปร | SS | df | MS |
|---|---|---|---|
| เนื่องจากความถดถอย (R) | $SSR = \Sigma(\hat{Y} - \bar{Y})^2$ | 1 | $MSR = \frac{SSR}{1}$ |
| ความคลาดเคลื่อน (E) | $SSE = \Sigma(Y - \hat{Y})^2$ | $n - 2$ | $MSE = \frac{SSE}{n-2}$ |
| รวม | $SST = \Sigma(Y - \bar{Y})^2$ | $n - 1$ | |

จะใช้แบบฝึกหัด 11.14 เป็นตัวอย่าง ข้อมูลคือ ค่าบำรุงรักษา และอายุการใช้งานของเครื่องจักร 6 เครื่อง

| เครื่องจักร | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| X = อายุ | 2 | 1 | 3 | 2 | 1 | 3 |
| Y = ค่าซ่อมแซม (1,000 บาท) | 70 | 40 | 100 | 80 | 30 | 100 |

$n = 6$ $\quad \Sigma X = 12, \Sigma X^2 = 28, \bar{X} = 2$

$\Sigma Y = 420, \Sigma Y^2 = 33800, \bar{Y} = 70$

$\Sigma XY = 970$

$SST = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 33{,}800 - (420)^2/6$

$= 4{,}400$ และมี $df = 6 - 1 = 5$

$$SSR = \frac{[(\Sigma(X - \bar{X})(Y - \bar{Y}))^2]}{\Sigma(X - \bar{X})^2}$$

$\Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma XY - (\Sigma X)(\Sigma Y)/n$

$= 970 - (12)(420)/6$

$= 130$

$\Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/6$

$= 4$

ดังนั้น SSR $= \frac{(130)^2}{4} = 4,225$ และมี df $= 1$

และ SSE $= SST - SSR$

$= 4,400 - 4,225$

$= 175$ และมี df $= 6 - 2 = 4$

| ที่มาของความผันแปร | SS | df | MS |
|---|---|---|---|
| ความถดถอย (R) | 4,225 | 1 | 4,225 |
| ความคลาดเคลื่อน (E) | 175 | 4 | 43.75 |
| รวม | 4,400 | 5 | |

จุดประสงค์หลักของการสร้างตารางวิเคราะห์ความแปรปรวนคือ การหาอัตราส่วน F หรือค่าสถิติ F

อัตราส่วน $F = \frac{MSR}{MSE} \sim f_{1, n-2, \alpha}$

จะเป็นตัวสถิติที่ใช้ทดสอบความลาดชัน หรือความสัมพันธ์เชิงเส้น

คือทดสอบ Ho : $\beta = 0$, Ha : $\beta \neq 0$

ซึ่งเดิมเราใช้ t - test คือ ตัวสถิติ

$$T = \frac{\hat{\beta}}{S_{\hat{\beta}}} = \frac{b}{S_b} \sim t_{(n-2), \alpha/2}$$

ตัวสถิติคู่นี้มีความสัมพันธ์กัน คือ

ค่าสถิติ $(T)^2 = F$

ค่าเปิดตาราง $(t_{(n-2,\alpha/2)})^2 = f_{1,(n-2),\alpha}$

1. Ho : $\beta = 0$ (อายุการใช้งาน และค่าซ่อมแซมไม่มีความสัมพันธ์เชิงเส้น)

2. Ha : $\beta \neq 0$ (อายุการใช้งานและค่าซ่อมแซมมีความสัมพันธ์เชิงเส้น)

3. ให้ $\alpha = .05$

4. จะปฏิเสธ Ho เมื่อ $F > f_{1,(6-2),.05} = 7.71$

5. $F = \frac{MSR}{MSE} = \frac{4225}{43.75} = 96.58$

6. $F = 96.57 > 7.71$ ตกอยู่ในเขตวิกฤต จึงปฏิเสธ Ho และสรุปว่าเส้นถดถอยมีความลาดชันไม่ใช่ 0 (มีความลาดชัน) นั่นคือ อายุการใช้งานและค่าซ่อมแซมมีความสัมพันธ์แบบเชิงเส้นตรง

**ถ้าใช้ t – test ทำดังนี้**

Ho : $\beta = 0$ , Ha : $\beta \neq 0$

$\alpha = .05$, จะปฏิเสธ Ho เมื่อ $T > t_{4,.025}$ หรือ

$T < -t_{4,.025}$ นั่นคือเมื่อ $|T| > 2.776$

$T = b/S_b$

$$b = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{130}{4} = 32.5$$

$$S_b = \sqrt{\frac{S_{y/x}}{\Sigma(X-\bar{X})^2}}$$

$$S^2_{y/x} = \frac{\Sigma(Y-\hat{Y})^2}{n-2} = \frac{SSE}{n-2} = MSE$$

ดังนั้น

$$S_b = \sqrt{\frac{MSE}{\Sigma(X-\bar{X})^2}} = \sqrt{\frac{43.75}{4}} = 3.307$$

ดังนั้น

$$T = \frac{32.5}{3.307} = 9.828$$

$T = 9.828 > 2.776$ ตกอยู่ในเขตวิกฤต จึงปฏิเสธ Ho และสรุปว่า X และ Y มีความสัมพันธ์เชิงเส้น เช่นเดียวกับการทดสอบแบบ F

และโปรดสังเกตว่า $(T)^2 = (9.828)^2 = F = 96.58$

ในขั้นนี้เราจึงสรุปได้ว่า ประโยชน์ของ ANOVA คือ

1. ใช้ทดสอบความสัมพันธ์เชิงเส้น
2. ได้ค่าประมาณของ $\sigma^2_{y/x}$ คือ $S^2_{y/x}$ = MSE

## 3. การวิเคราะห์สหสัมพันธ์

การวิเคราะห์สหสัมพันธ์เป็นเครื่องมือทางสถิติ เพื่อใช้อธิบายขนาด (degree) ของความสัมพันธ์เชิงเส้นระหว่างตัวแปรคู่หนึ่ง เรามักทำการวิเคราะห์สหสัมพันธ์ควบคู่กับการวิเคราะห์ความถดถอย เพื่อวัดว่า เส้นถดถอยอธิบายความผันแปรของตัวแปรพึ่งพา Y ได้ดีเพียงไร เราอาจศึกษาเฉพาะการวิเคราะห์สหสัมพันธ์เพื่อวัดดีกรีความเกี่ยวพันกัน (association) ระหว่างตัวแปรคู่หนึ่ง

นักสถิติได้สร้างเครื่องมือที่ใช้วัดสหสัมพันธ์ระหว่างตัวแปร 2 ตัวไว้ 2 อย่าง คือ **สัมประสิทธิ์การตัดสินใจ** (coefficient of determination) และ**สัมประสิทธิ์สหสัมพันธ์** (coefficient of correlation)

### สัมประสิทธิ์การตัดสินใจ

สัมประสิทธิ์การตัดสินใจ คือ มาตราเบื้องต้นที่ใช้วัดความแรงของการเกี่ยวพันกันระหว่างตัวแปรคู่หนึ่ง คือ X และ Y แต่เนื่องจากเราใช้ข้อมูลจากตัวอย่างเพื่อสร้างเส้นถดถอย เราจึงเรียกมาตรานี้ว่า **สัมประสิทธิ์การตัดสินใจของตัวอย่าง** (sample coeffiecient of determination) ใช้สัญลักษณ์ว่า $r^2$ และเป็นค่าประมาณของพารามิเตอร์ $\rho^2$ ค่า $r^2$ มาจากความสัมพันธ์ระหว่างความผันแปรของค่า Y ต่าง ๆ จากตัวอย่างที่เก็บมา 2 ชนิด คือ

1. ความผันแปรของ Y โดยรอบเส้นถดถอย
2. ความผันแปรของ Y โดยรอบค่าเฉลี่ยของตัวเอง

คำว่า "ความผันแปร" ในทางสถิติเราหมายถึง "ผลรวมของกำลังสองของค่าเบี่ยงเบน" ดังนั้น เราจะหาความผันแปรทั้ง 2 ชนิดได้ดังนี้

ความผันแปรของ Y โดยรอบเส้นถดถอย $= \Sigma(Y - \hat{Y})^2$ (11.16)

และ

ความผันแปรของ Y โดยรอบค่าเฉลี่ยของตัวเอง $= \Sigma(Y - \bar{Y})^2$ (11.17)

เมื่อนำอัตราส่วนของความผันแปรคู่นี้หักออกจาก 1 จะได้สัมประสิทธิ์การตัดสินใจของตัวอย่าง ซึ่งใช้สัญลักษณ์ $r^2$:

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2} \qquad (11.18)$$

ยังมีสูตรอื่นที่ใช้คำนวณค่า $r^2$ ซึ่งจะได้กล่าวถึงภายหลัง

## การอธิบายค่า $r^2$

ก่อนอื่นเราจะพิจารณาค่าต่ำสุดและสูงสุดของ $r^2$ ซึ่งจะเกิดขึ้นเมื่อตัวแปร X และ Y มีค่าในตารางที่ 11.5 และ 11.6

ตารางที่ 5 แสดงสหสัมพันธ์แบบสมบูรณ์ระหว่างตัวแปร 2 ตัว

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\Sigma X = 36$ |
| Y | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 32 | $\Sigma Y = 144$ |

$$\bar{Y} = \frac{144}{8} = 18 \,,\ \bar{X} = \frac{36}{8} = 4.5$$

จะเห็นว่าค่าของ X เพิ่มขึ้นทีละ 1 หน่วย แต่ค่า Y เพิ่มทีละ 4 หน่วย ดังนั้น เส้นถดถอยจะมีความลาดชัน = 4 หรือ b = 4

ส่วน $a = \bar{Y} - b\bar{X}$

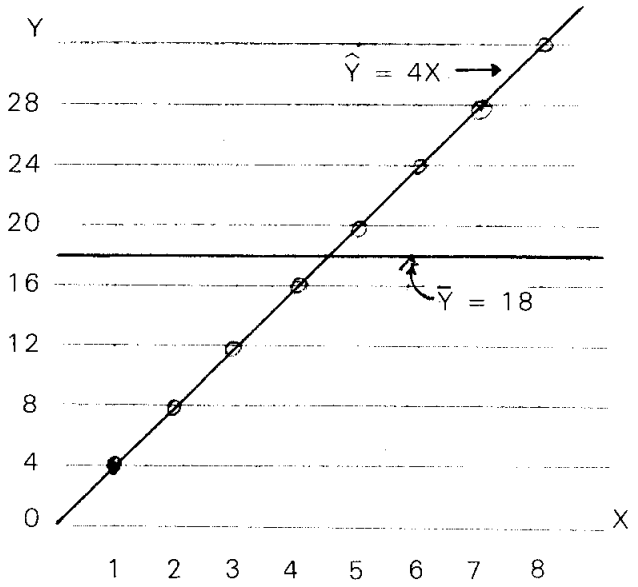
$= 18 - 4(4.5)$

$= 0$

แสดงว่า เส้นถดถอยผ่านจุดกำเนิด ดังนั้นสมการถดถอยคือ

$$\hat{Y} = 4X$$

และมีกราฟเส้นตรงในรูป 11.14




**รูปที่** 11.14
แสดงสหสัมพันธ์แบบสมบูรณ์ระหว่าง X และ Y จะเห็นว่าทุกจุดอยู่บนเส้นถดถอย

เราจะคำนวณค่า $r^2$ ดังนี้

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

แต่ $\Sigma(Y - \hat{Y})^2 = 0$ เพราะทุกจุดอยู่บนเส้นถดถอย นั่นคือ $Y = \hat{Y}$

| | | | | |
|---|---|---|---|---|
| และ $\Sigma(Y - \bar{Y})^2$ | $= ( 4 - 18)^2$ | $= (-14)^2$ | = | 196 |
| | $+ ( 8 - 18)^2$ | $= (-10)^2$ | = | 100 |
| | $+ (12 - 18)^2$ | $= (-6)^2$ | = | 36 |
| | $+ (16 - 18)^2$ | $= (-2)^2$ | = | 4 |
| | $+ (20 - 18)^2$ | $= (2)^2$ | = | 4 |
| | $+ (24 - 18)^2$ | $= (6)^2$ | = | 36 |
| | $+ (28 - 18)^2$ | $= (10)^2$ | = | 100 |
| | $+ (32 - 18)$ | $= (14)^2$ | = | 196 |
| | | $\Sigma(Y - \bar{Y})^2$ | ⟶ | 672 |

แทนค่าในสูตรจะได้

$$r^2 = 1 - \frac{0}{672}$$

$= 1$ ⟵ สัมประสิทธิ์การตัดสินใจของตัวอย่าง

กรณีที่ตัวแปรคู่นั้นมีสหสัมพันธ์แบบสมบูรณ์

ในกรณีที่ $r^2 = 1$ แสดงว่าเส้นถดถอยเป็นตัวประมาณค่าที่สมบูรณ์ เพราะไม่มีความผิดพลาดในการประมาณค่า Y ด้วย $\hat{Y}$ เลย

**ตารางที่** 11.6 แสดงค่าของ X และ Y เมื่อมีความสัมพันธ์ในแนวขนานกับค่าแกน X นั่นคือ Y มีค่าคงเดิม เมื่อ X เปลี่ยนแปลง ทำให้มีสหสัมพันธ์เป็น 0 คือไม่มีสหสัมพันธ์กันโดยสิ้นเชิง

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\Sigma X = 36$ |
| Y | 6 | 12 | 6 | 12 | 6 | 12 | 6 | 12 | $\Sigma Y = 72$ |

$\bar{X} = \frac{36}{8} = 4.5, \quad \bar{Y} = \frac{72}{8} = 9$

กรณีนี้เส้นถดถอยจะไม่มีความลาดชัน หรือ $b = 0$ เพราะเมื่อ X เพิ่มทีละ 1 หน่วย Y จะเพิ่ม 6 หน่วย และลด 6 หน่วย เท่ากับอัตราการเปลี่ยนแปลงเป็น 0 จะเห็นว่าเมื่อ X เปลี่ยนไปทีละหน่วย Y ยังคงมีค่าไม่เปลี่ยนแปลง

ดังนั้น $a = \bar{Y} - b\bar{X}$

$= 9 - 0(4.5)$

$= 9 \quad \longleftarrow \bar{Y}$

และเส้นถดถอย คือ

$\hat{Y} = 9$ หรือ $\hat{Y} = \bar{Y}$

เมื่อหาความผันแปรโดยรอบเส้นถดถอยคือ $\Sigma(Y - \hat{Y})^2$ ได้ดังนี้

$$\Sigma(Y - \hat{Y})^2 = (6 - 9)^2 = (-3)^2 = 9$$
$$+ (12 - 9)^2 = (3)^2 = 9$$
$$+ (6 - 9)^2 = (-3)^2 = 9$$
$$+ (12 - 9)^2 = (3)^2 = 9$$
$$+ (6 - 9)^2 = (-3)^2 = 9$$
$$+ (12 - 9)^2 = (3)^2 = 9$$

$$+ (6 - 9)^2 = (-3)^2 = 9$$
$$+ (12 - 9)^2 = (3)^2 = 9$$
$$72 \longleftarrow \Sigma(Y - \hat{Y})^2$$

และจะหาความผันแปรของ Y โดยรอบค่าเฉลี่ยของตัวเอง คือ $\Sigma(Y - \bar{Y})^2$ ย่อมจะได้ 72 เท่ากับ $\Sigma(Y - \hat{Y})^2$ เพราะ $\hat{Y} = \bar{Y}$

$$\begin{aligned}
\Sigma(Y - \hat{Y})^2 &= (6 - 9)^2 = (-3)^2 = 9 \\
&+ (12 - 9)^2 = (3)^2 = 9 \\
&+ (6 - 9)^2 = (-3)^2 = 9 \\
&+ (12 - 9)^2 = (3)^2 = 9 \\
&+ (6 - 9)^2 = (-3)^2 = 9 \\
&+ (12 - 9)^2 = (3)^2 = 9 \\
&+ (6 - 9)^2 = (-3)^2 = 9 \\
&+ (12.9)^2 = (3)^2 = 9 \\
&\qquad 72 \longleftarrow \Sigma(Y - \bar{Y})^2
\end{aligned}$$

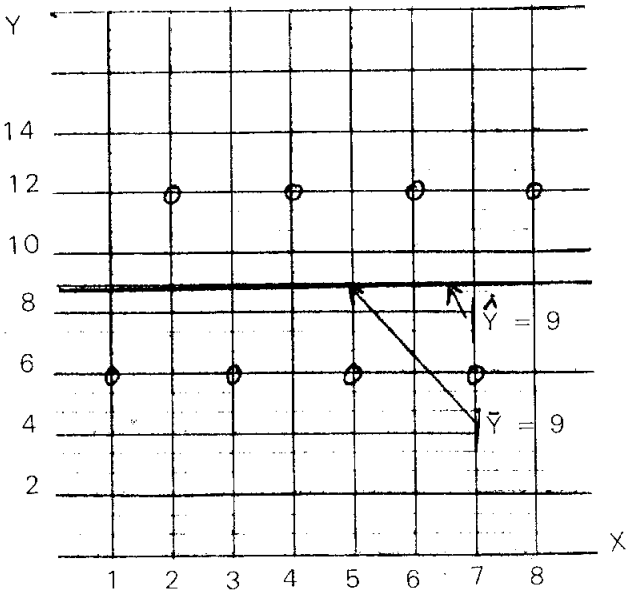
แทนค่าในสูตรของ $r^2$ จะได้

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

$$= 1 - \frac{72}{72}$$

$$= 1 - 1$$
$$= 0$$

จึงสรุปได้ว่า $r^2 = 0$ เมื่อตัวแปรคู่นั้น
ไม่มีสหสัมพันธ์กัน



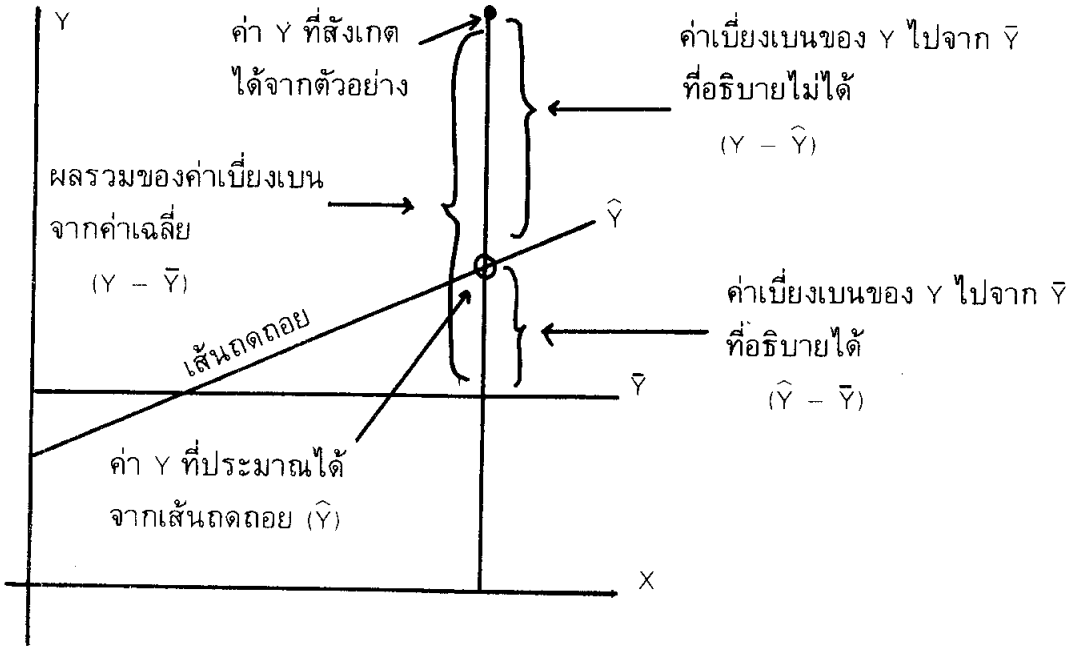

**รูปที่** 11.15 แสดงสหสัมพันธ์ระหว่าง X และ Y = 0 เพราะค่า Y เป็นค่าซ้ำ ๆ กัน ในขณะที่ X เปลี่ยนแปลง

แต่ในทางปฏิบัติเรามักพบปัญหาว่า $r^2$ ไม่เป็น 0 หรือ 1 ทีเดียว แต่จะมีค่าอยู่ระหว่าง 2 ค่านี้ หากค่า $r^2$ มีค่าใกล้ 1 จะแสดงว่าตัวแปรคู่นั้นมีสหสัมพันธ์ค่อนข้างสูง แต่ถ้าค่า $r^2$ มีค่าใกล้ 0 แสดงว่าตัวแปรคู่นั้นมีสหสัมพันธ์ค่อนข้างน้อย

จุดสำคัญ คือต้องแยกให้ออกว่า $r^2$ เป็นมาตราวัดความแรงของความสัมพันธ์**เชิงเส้น** ระหว่าง 2 ตัวแปร ตัวอย่างเช่น ถ้าเรามีจุด (X, Y) จำนวนมากเรียงกันในลักษณะเส้นรอบวงกลม (แต่อยู่ในลักษณะกระจายแบบสุ่ม) จะเห็นได้ชัดเจนว่า จุดเหล่านั้น มีความสัมพันธ์กันแน่นอน (เพราะอยู่ในวงกลมเดียวกัน) แต่ถ้าหาค่า $r^2$ จะมีค่าใกล้ 0 ทั้งนี้ เพราะจุดเหล่านั้น ไม่ได้**เรียงกันในลักษณะเชิงเส้นตรง**

### ความหมายของ $r^2$ อีกอย่างหนึ่ง

นักสถิติยังอธิบายความหมายของ $r^2$ อีกทางหนึ่ง คือ **หมายถึงจำนวนความผันแปรของ Y ที่อธิบายได้โดยเส้นถดถอย** การอธิบายความหมายนี้ควรดูรูป 11.16 ประกอบ




**รูปที่** 11.16 แสดงค่าเบี่ยงเบนทั้งหมด, ค่าเบี่ยงเบนที่อธิบายได้ ค่าเบี่ยงเบนที่อธิบายไม่ได้ของค่าสังเกต Y เพียง 1 ตัว

จากรูปที่ 11.16 ค่าสังเกต Y คือวงกลมสีดำ ถ้าเราใช้ค่าเฉลี่ย $\bar{Y}$ เป็นค่าประมาณของ Y เราจะได้ค่าเบี่ยงเบนทั้งหมดคือระยะทาง $(Y - \bar{Y})$ **แต่ถ้าใช้เส้นถดถอย** $\hat{Y}$ เป็นค่าประมาณของ Y เราจะได้ค่าประมาณที่ดีขึ้น อย่างไรก็ตามเส้นถดถอยจะอธิบายได้เพียงส่วนหนึ่งของค่าเบี่ยงเบนทั้งหมดคือ ระยะทาง $(\hat{Y} - \bar{Y})$ ส่วนที่เหลือจากค่าเบี่ยงเบนทั้งหมดคือระยะทาง $(Y - \hat{Y})$ ยังคงอธิบายไม่ได้

ถ้าเราหาค่าเบี่ยงเบนทั้งหลายนี้ จากทุก ๆ ค่าสังเกต Y ที่เก็บจากตัวอย่าง n ตัวนั้น และหาในรูปผลบวกกำลังสอง เราจะได้ผลบวกกำลังสองของค่าเบี่ยงเบนทั้งหมด จากทุก ๆ จุดไปจากค่าเฉลี่ยของตัวเอง คือ

total sum of squares $= \Sigma(Y - \bar{Y})^2$ หรือ SST (11.19)

และส่วนของความผันแปรทั้งหมดที่อธิบายได้ คือ

regression sum of squares $= \Sigma(\hat{Y} - \bar{Y})^2$ หรือ SSR (11.20).

และส่วนของความผันแปรทั้งหมดที่อธิบายไม่ได้ คือ

error sum of sguares $= \Sigma(Y - \hat{Y})^2$ หรือ SSE (11.21)

ดังนั้น $\frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$ คือส่วนของความผันแปรทั้งหมดที่อธิบายสาเหตุไม่ได้

และ $\frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$ คือส่วนของความผันแปรทั้งหมดที่อธิบายได้

ดังนั้น
$$r^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} \qquad (11.22)$$

และ
$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2} \qquad (11.23)$$

สูตรที่ (11.22) เป็นสูตรใหม่ สูตร (11.23) ตรงกับสูตรเดิมคือสูตร (11.18) เพราะ $r^2$ วัดดีกรีของความเกี่ยวพันกันระหว่าง X และ Y

สูตรที่ (11.22) และ (11.23) อาจไม่สะดวกในการคำนวณ ยังมีอีกสูตรที่ง่ายต่อการคำนวณ คือสูตรที่ (11.24) คือ

$$r^2 = \frac{a\Sigma Y + b\Sigma XY - n\bar{Y}^2}{\Sigma Y^2 - n\bar{Y}^2} \qquad (11.24)$$

ในเมื่อ

$r^2$ = สัมประสิทธิ์การตัดสินใจจากตัวอย่าง
a = จุดตัดบนแกน Y
b = ความลาดชันของเส้นถดถอย
n = ขนาดตัวอย่าง = จำนวนจุด
X = ค่าของตัวแปรอิสระ
Y = ค่าของตัวแปรพึ่งพา
$\bar{Y}$ = ค่าเฉลี่ยของตัวแปรพึ่งพา

และถ้าหากเราสร้างตารางวิเคราะห์ความแปรปรวนของความถดถอยคือตารางที่ 11.4 เราจะหาค่า $r^2$ ได้อีกวิธีหนึ่ง คือ

$$r^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} = \frac{SSR}{SST} \qquad (11.25)$$

**ตารางที่** 11.7 แสดงการหาค่า $r^2$ โดยสูตรต่าง ๆ

| ปี<br>n = 6 | ค่าใช้จ่ายในการวิจัยพัฒนา X | ผลกำไรต่อปี Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| | (2) | (3) | (2)x(3) | (2) | (3) |
| (1) | | | | | |
| 2525 | 5 | 31 | 155 | 25 | 961 |
| 2524 | 11 | 40 | 440 | 121 | 1,600 |
| 2523 | 4 | 30 | 120 | 16 | 900 |
| 2522 | 5 | 34 | 170 | 25 | 1,156 |
| 2521 | 3 | 25 | 75 | 9 | 625 |
| 2520 | 2 | 20 | 40 | 4 | 400 |
| | $\Sigma X = 30$ | $\Sigma Y = 180$ | $\Sigma XY = 1,000$ | $\Sigma X^2 = 200$ | $\Sigma Y^2 = 5,642$ |
| | $\bar{X} = \frac{30}{6}$ | $\bar{Y} = \frac{180}{6}$ | | | |
| | $= 5$ | $= 30$ | | | |

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} = \frac{1000 - 6(5)(30)}{200 - 6(5)^2}$$

$$= \frac{1000 - 900}{200 - 150}$$

$$= \frac{100}{50}$$

$$= 2$$

$$a = \bar{Y} - b\bar{X}$$

$$= 30 - 2(5)$$

$$= 20$$

สมการประมาณค่า คือ

$$\hat{Y} = 20 + 2X$$

| X | Y | $\hat{Y}$ | $(Y - \hat{Y})^2$ | $(\hat{Y} - \bar{Y})^2$ | $(Y - \bar{Y})^2$ |
|---|---|---|---|---|---|
| 5 | 31 | 30 | 1 | 0 | 1 |
| 11 | 40 | 42 | 4 | 144 | 100 |
| 4 | 30 | 28 | 4 | 4 | 0 |
| 5 | 34 | 30 | 16 | 0 | 16 |
| 3 | 25 | 26 | 1 | 16 | 25 |
| 2 | 20 | 24 | 16 | 36 | 100 |
| | | | 42 | 200 | 242 |

ถ้าคำนวณ $r^2$ โดยสูตร (11.22)

$$r^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} = \frac{200}{242} = .826$$

ถ้าใช้สูตร (11.23)

$$r^2 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

$$= 1 - \frac{42}{242}$$

$$= 1 - .174 = .826$$

ถ้าใช้สูตร (11.24)

$$r^2 = \frac{a\Sigma Y = b\Sigma XY - n\bar{Y}^2}{\Sigma Y^2 - n\bar{Y}^2}$$

$$= \frac{20(180) + 2(1000) - 6(30)^2}{5642 - 6(30)^2}$$

$$= \frac{3600 + 2000 - 5400}{5642 - 5400} = \frac{200}{242} = .826$$

ANOVA

| ที่มา | SS | df | MS |
|---|---|---|---|
| ความถดถอย (R) | 200 | 1 | 200 |
| ความคลาดเคลื่อน (E) | 42 | 4 | 10.5 |
| รวม | 242 | 5 | |

$$R^2 = \frac{SSR}{SST} = \frac{200}{242} = .826$$

สรุปได้ว่า ความผันแปรของค่าใช้จ่ายวิจัยและพัฒนา (ตัวแปรอิสระ X) อธิบายได้ 82.6% ของความผันแปรของผลกำไรต่อปี (ตัวแปรพึ่งพา Y)

## สัมประสิทธิ์สหสัมพันธ์

(The coefficient of correlation)

สัมประสิทธิ์แห่งสหสัมพันธ์เป็นเครื่องมืออีกอันหนึ่งที่ชี้ให้เห็นว่าตัวแปรตัวหนึ่งใช้อธิบายตัวแปรอีกตัวหนึ่งได้ดีมากน้อยเพียงไร ค่าที่คำนวณได้จากตัวอย่างเรียกว่า สัมประสิทธิ์แห่งสหสัมพันธ์จากตัวอย่าง และใช้ r เป็นค่าสถิติที่ใช้ประมาณพารามิเตอร์ $\rho$ มีวิธีหาค่า r 2 วิธี คือ

1. หาจากสูตรโดยตรงคือ

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

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

2. หาจากสัมประสิทธิ์การตัดสินใจ

โดยที่ $r = \sqrt{r^2}$

และ r จะมีเครื่องหมายเหมือนความลาดชันของเส้นถดถอย (b) นั่นคือ ถ้า X และ Y มีความสัมพันธ์ใน

ทางกลับกัน r จะมีเครื่องหมายลบ และจะมีค่าอยู่ระหว่าง 0 ถึง -1 แต่ถ้า X และ Y มีความสัมพันธ์ในทางเดียวกัน r จะมีเครื่องหมายบวก และจะมีค่าอยู่ระหว่าง 0 ถึง 1 ดังแสดงในรูปที่ 11.17

การอธิบายความหมายของสัมประสิทธิ์สหสัมพันธ์ยากกว่า การอธิบายความหมายของ $r^2$ สมมุติคำนวณได้ r = .9 จะหมายความว่าอย่างไร? อย่าลืมว่า เมื่อ r = .9 ก็คือ $r^2$ = .81 นั่นเอง ซึ่งหมายความว่า 81% ของความผันแปรทั้งหมดของ Y สามารถอธิบายได้โดยเส้นถดถอย (หรือตัวแปรอิสระ X) และเราทราบว่า $r = \sqrt{r^2}$ และค่าสูงสุดของ r = 1 เมื่อ r = .9 ก็แสดงว่า X และ Y มีสหสัมพันธ์ในทิศทางเดียวกันค่อนข้างสูง เราก็คงจะอธิบายค่าของ r ได้เพียงแค่นี้เอง

จากข้อมูลในตารางที่ 11.7 เรื่องความสัมพันธ์ระหว่างรายจ่ายเพื่อพัฒนาและวิจัยกับผลกำไรต่อปี ซึ่งคำนวณค่า $r^2$ = .826

ดังนั้น $r = \sqrt{r^2} = \sqrt{.826}$

$= .909$

r จะมีเครื่องหมายเป็นบวก เพราะความสัมพันธ์ระหว่างตัวแปรคู่นี้เป็นไปในทิศทางเดียวกัน (b มีค่าเป็นบวก)

ถ้าหาจากสูตรโดยตรง โดยไม่หาจาก $r^2$

$$r = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\sqrt{(\Sigma X^2 - n\bar{X}^2)(\Sigma Y^2 - n\bar{Y}^2)}}$$

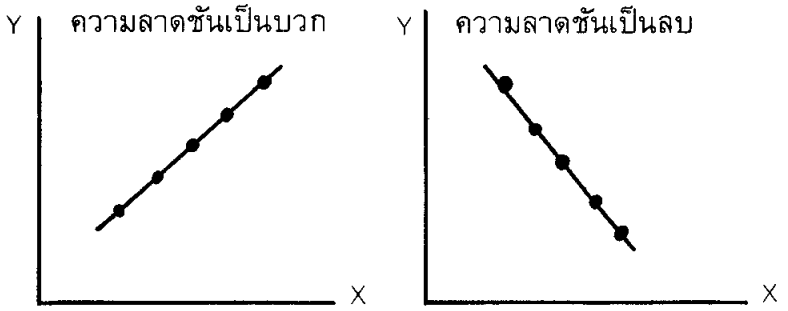
$$= \frac{1000 - 6(5)(30)}{\sqrt{(200 - 6(5)^2)(5642 - 6(30^2)}}$$

$$= \frac{100}{\sqrt{(50)(242)}}$$

$$= \frac{100}{110}$$

$$= 0.909$$

(ก) $r^2 = 1$ และ $r = 1$ (ข) $r^2 = 1$ และ $r = -1$




**รูปที่** 11.17
แสดงค่าต่าง ๆ
ของ r

(ค) $r^2 = .81$ และ $r = .9$ (ง) $r^2 = .49$ และ $r = -.7$ (ง) $r^2 = 0$ และ $r = 0$


**การทดสอบสมมติฐาน**

อาจมีบ่อยครั้งที่เราต้องการทราบว่า ค่า r ที่คำนวณได้จากตัวอย่างมีค่าโตพอที่จะแสดงว่า ตัวแปรคู่นั้นมีสหสัมพันธ์ในประชากรด้วยหรือไม่มี นั่นคือ เราอาจต้องการทดสอบค่าของ $\rho$ ว่าเป็น 0 หรือไม่

ก่อนการทดสอบต้องมีเงื่อนไขว่า ตัวแปรทั้งคู่นั้นต่างก็มีการแจกแจงแบบปกติ เพื่อเราจะได้ใช้การทดสอบแบบ t ได้ และจะมีขั้นตอนการทดสอบดังนี้

1. Ho : $\rho = 0$ (ตัวแปรคู่นั้นไม่มีสหสัมพันธ์ในประชากร)
2. Ha : $\rho \neq 0$ (ตัวแปรคู่นั้นมีสหสัมพันธ์ในประชากร)

หรือ Ha : $\rho > 0$ (ตัวแปรคู่นั้นมีสหสัมพันธ์กัน และเป็นแบบเชิงบวก)

หรือ Ha : $\rho < 0$ (ตัวแปรคู่นั้นมีสหสัมพันธ์กัน และเป็นแบบตรงข้าม)

3. กำหนดระดับนัยสำคัญ $\alpha$
4. จะปฏิเสธ Ho เมื่อ $|T| > t_{\alpha/2,\ \nu}$ สำหรับ Ha : $\rho \neq 0$

เมื่อ $T > t_{\alpha, \nu}$ สำหรับ Ha : $\rho > 0$

เมื่อ $T < -t_{\alpha, \nu}$ สำหรับ Ha . $\rho < 0$

โดยที่ $\nu = n - 2$

5. ตัวสถิติที่ใช้ทดสอบคือ

$$T = r\sqrt{\frac{n-2}{1-r^2}}$$

และ T จะมีการแจกแจงแบบค่า t ในตารางด้วย $\nu = n - 2$

6. ปฏิเสธ Ho ถ้าค่า T ตกอยู่ในเขตวิกฤต

โปรดสังเกตว่า ค่าสถิติ $T = r\sqrt{\frac{n-2}{1-r}2}$ จะมีค่าโตเมื่อ r เป็นค่าโต (คิด absolute value) ดังนั้น เมื่อ T มีค่าโตมากจนตกในเขตวิกฤต เราจะปฏิเสธ Ho และแสดงว่ามีสหสัมพันธ์ระหว่างตัวแปรคู่นั้น

จากตัวอย่างในตาราง 11.7 ซึ่งคำนวณค่า r = .909 และ $r^2$ = .826 ถ้าเราต้องการทดสอบว่าค่าใช้จ่ายในการพัฒนาและวิจัย มีสหสัมพันธ์กับรายได้จากการขายแบบเชิงบวก (คือทิศทางเดียวกัน) มีวิธีการดังนี้

1. Ho : $\rho = 0$

2. Ha : $\rho > 0$

3. $\alpha = .05$

4. จะปฏิเสธ Ho เมื่อ $T > t_{.05, 4} = 2.132$

5.

$$T = r\sqrt{\frac{n-2}{1-r}2}$$

$$= .909\sqrt{\frac{6-2}{1-.826}}$$

$$= .090\sqrt{22.988} \qquad = .090(4.795)$$

$$= 4.36$$

6. ค่า T = 4.36 สูงกว่า 2.132 จึงอยู่ในขตวิกฤต เราจึงปฏิเสธ Ho และสรุปว่า ค่าใช้จ่ายในการพัฒนาและวิจัยกับรายได้จากการขายต่อปี มีสหสัมพันธ์แบบเชิงบวก

พึงสังเกตว่าค่า T มีคำนวณได้ จะเท่ากับค่า T จากการทดสอบ Ho : $\beta = 0$ จาก ANOVA

$$F = \frac{MSR}{MSE} = \frac{200}{10.5} = 19.05$$

$$\sqrt{F} = \sqrt{19.5} = 4.36 = T$$

โดยที่

$$T = \frac{b}{S_b} = \frac{2}{.4582} = 4.36$$

$$S_b = \sqrt{\frac{MSE}{\Sigma X^2 - n\bar{X}^2}}$$

$$= \sqrt{\frac{10.5}{50}}$$

$$= \sqrt{.21}$$

$$= .4582$$

ดังนั้น ความหมายของการทดสอบสมมุติฐาน

Ho : $\rho = 0$ และ Ho : $\beta = 0$

จึงเหมือนกัน

## ข้อควรระวังในการอธิบายค่า r

1. ค่า $r^2$ จะบอกความแรงของขนาดความสัมพันธ์เชิงเส้นได้ดีกว่าค่า r เช่นเมื่อ r = .5, $r^2$ จะมีค่าเพียง .25 และถ้าจะให้ค่า $r^2 > .5$ r จะต้องมีค่า .7 ขึ้นไป หรือเมื่อ r = 0.3, $r^2$ = .09 ซึ่งแสดงว่า X อธิบายความผันแปรของ Y ได้เพียง 9% ส่วนที่เหลืออีก 91% ต้องใช้ตัวแปรอื่น ๆ ช่วยอธิบาย
2. $r^2$ **ไม่ได้วัดความลาดชันของเส้นถดถอย** เช่นถ้าค่า $r^2$ สูง ($r^2 \to 1$) ก็ไม่จำเป็นที่ $\beta$ หรือ b จะต้องมีค่าสูงด้วย เช่นในรูปที่ 11.18 ซึ่ง $r^2$ มีค่าเป็น 1 ทั้ง 2 รูป แต่เส้นทั้ง 2 มีความลาดชันต่างกัน รูป (ก) จะมีความลาดชันสูงกว่า ทั้งนี้จะอธิบายได้ ดังนี้

   เมื่อ $r^2 = 1$

   คือ $\frac{SSR}{SST} = 1$

หรือ $\frac{b^2\Sigma(X - \bar{X})^2}{\Sigma(Y - \bar{Y})^2} = 1$

หรือ $\frac{b^2\Sigma(X - \bar{X})^2/(n - 1)}{\Sigma(Y - \bar{Y})^2/(n - 1)} = 1$

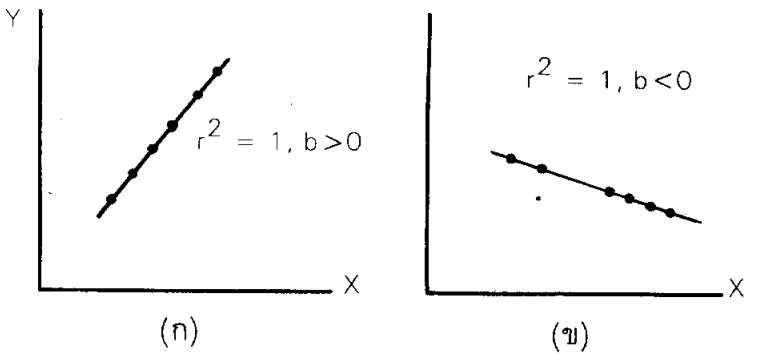
หรือ $\frac{b^2 S_X^2}{S_Y^2} = 1$

ดังนั้น $b^2 = \frac{S_Y^2}{S_X^2}$ (เมื่อ $r^2 = 1$)

หมายความว่า แม้ว่าข้อมูล 2 ชุดใด ๆ จะมีความผันแปรของ X เท่ากัน แต่ถ้าความผันแปรของ Y ของกลุ่มใดสูงกว่า จะทำให้ค่า b สูงกว่าด้วย รูป (ก) มีความผันแปรของ Y สูงกว่ารูป (ข) ความลาดชันจึงมากกว่า

---

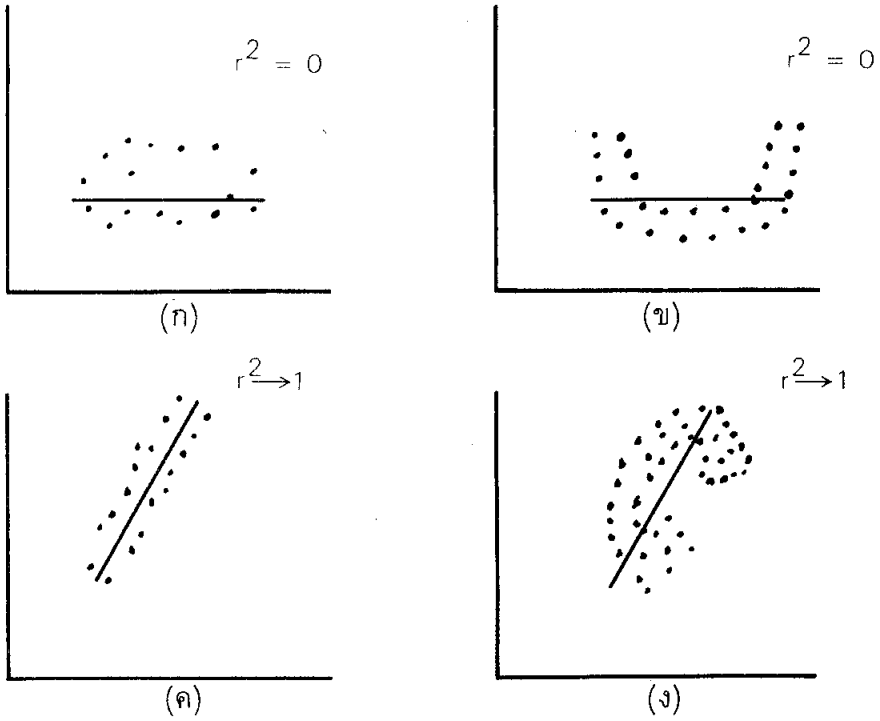
**รูปที่** 11.18
แสดงความสัมพันธ์เชิงเส้นแบบสมบูรณ์




3. $r^2$ **ไม่ได้วัดความเหมาะสมของตัวแบบเชิงเส้น** เช่น ในรูปที่ 11.19 รูป (ก) และ (ข) มีค่า $r^2 = 0$ เหมือนกัน รูป (ก) แสดงว่า X และ Y ไม่มีลักษณะความสัมพันธ์แบบใดทั้งสิ้น แต่รูป (ข) แสดงหลักฐานอย่างมั่นคงว่า X และ Y มีลักษณะความสัมพันธ์แบบเส้นโค้ง (nonlinear association)

**ส่วนในรูป** (ค) และ (ง) $r^2$ **ที่คำนวณจะมีค่าสูง** และเส้นตรงจะสอดคล้องกับข้อมูลได้อย่างดีในรูป (ค) แต่ในรูป (ง) ตัวแบบเชิงเส้นกลับไม่เหมาะสมกับข้อมูล

รูปที่ 11.19 แสดงว่าค่า $r^2$ ไม่ได้ใช้วัดความเหมาะสมของตัวแบบเชิงเส้น




## แบบฝึกหัดทบทวน

11.17 ข้อมูลข้างล่างคือรายได้ต่อครอบครัว และค่าใช้จ่ายสำหรับอาหารของครอบครัวที่สุ่มมา 8 ครอบครัว คือ

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| รายได้ (1,000 บาท) | 8 | 12 | 9 | 24 | 13 | 37 | 19 | 16 |
| เปอร์เซ็นต์รายจ่ายสำหรับอาหาร | 36 | 25 | 33 | 15 | 28 | 19 | 20 | 22 |

ก) จงสร้างสมการประมาณค่าที่ดีที่สุดที่ใช้อธิบายความสัมพันธ์ของข้อมูล

(Y = 34.77 − 0.5809X)

ข) จงหา $S_{y/x}$ (4.963)

ค) จงหาช่วงเชื่อมั่น 90% ของเปอร์เซ็นต์รายจ่ายสำหรับอาหารของครอบครัวที่มีรายได้ 25,000 บาท (15.7375, 24.7575)

11.18 ข้อมูลข้างล่างคือ จำนวนขายเป็นพันบาทต่อเดือน และค่าประมาณของจำนวนขาย เป็นเวลา 10 เดือน ของบริษัทขายเครื่องกีฬา จงหาสัมประสิทธิ์การตัดสินใจจากตัวอย่าง และสัมประสิทธิ์สหสัมพันธ์จากตัวอย่าง

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 55 | 64 | 54 | 63 | 68 | 70 | 76 | 66 | 75 | 74 |
| $\hat{Y}$ | 55.5 | 59.5 | 60.5 | 63.5 | 67.5 | 65.5 | 73.5 | 70.5 | 72.5 | 76.5 |

($r = .87991$, $r^2 = .7733$)

11.19 ผู้จัดการบริษัทขายน้ำกรองสังเกตว่า ผู้ใช้น้ำของบริษัทมีการเปลี่ยนแปลงจำนวนบริโภคน้ำ ซึ่งแต่เดิมเขาพบว่า เส้นแสดงความสัมพันธ์ระหว่างจำนวนลูกค้า (บ้าน) และจำนวนบริโภค มีความลาดชัน 13 หน่วย จงใช้ระดับนัยสำคัญ .10 ทดสอบว่า ความลาดชันเพิ่มขึ้นหรือไม่ จากข้อมูลต่อไปนี้

| | | | | | | |
|---|---|---|---|---|---|---|
| จำนวนบ้าน (1,000 หลัง) | 8.1 | 7.8 | 8.4 | 7.6 | 8.0 | 8.1 |
| จำนวนการบริโภคน้ำ (10,000 แกลลอน) | 94 | 83 | 97 | 85 | 89 | 92 |

($b = 17.894736$, $T = 1.277$, ยอมรับ Ho ว่าความลาดชันไม่เปลี่ยนแปลง)

11.20 ในการสำรวจตลาด 10 ท้องที่ของบริษัทจำหน่ายรถจักรยานยนต์พบว่า จำนวนเงินที่ใช้โฆษณาและจำนวนขายจักรยานมีความสัมพันธ์แบบเชิงเส้นโดยมีความลาดชันเป็น 1.5 หน่วย และความคลาดเคลื่อนมาตรฐานของความลาดชันเป็น .35 ข้อมูลนี้จะขัดแย้งที่ระดับนัยสำคัญ .10 กับคำกล่าวอ้างของฝ่ายตลาดที่ว่า จะขายมอเตอร์ไซค์ได้เพิ่มขึ้น 60,000 คัน สำหรับรายจ่ายค่าโฆษณาที่เพิ่มขึ้น 25 ล้านบาท ($T = -2.57$, ปฏิเสธ Ho, จำนวนขายเพิ่มขึ้นต่ำกว่า 60,000 คัน)

11.21 ข้อใดที่เป็นความแตกต่างระหว่างสัมประสิทธิ์การตัดสินใจ และสัมประสิทธิ์สหสัมพันธ์ (ให้เลือกเพียงข้อเดียว)

ก) บอกได้ว่า ความลาดชันของเส้นถดถอยเป็นบวกหรือลบ

ข) วัดความแรงของความเกี่ยวพันระหว่าง 2 ตัวแปร

ค) จะไม่มีค่าเกิน 1

ง) วัดเปอร์เซ็นต์ของความผันแปรที่อธิบายโดยเส้นถดถอย

11.22 โรงงานผลิตเครื่องปรับอากาศทราบว่า จำนวนขายในฤดูร้อนจะไม่แน่นอน และมีความสัมพันธ์กับอุณหภูมิ จากการศึกษามา 6 ปี ได้ข้อมูลดังนี้

| อุณหภูมิ (ซ) | 29° | 33° | 35° | 34° | 36° | 31° |
|---|---|---|---|---|---|---|
| จำนวนขาย (เครื่อง) | 66 | 74 | 72 | 76 | 78 | 72 |

จงสร้างสมการที่อธิบายความสัมพันธ์ของข้อมูลนี้ ได้ดีที่สุด

(Y = 28.35 + 1.35X)

11.23 ในการวิเคราะห์ความถดถอย เราสามารถกำหนดค่า X ได้ล่วงหน้าแต่ค่า Y เป็นตัวแปรเชิงสุ่ม อยากทราบว่า ค่าของ X และ Y จะเป็นแบบใดเมื่อเราวิเคราะห์สหสัมพันธ์

11.24 ถ้าเรามีตัวอย่าง 2 ชุด และมีค่า r = 0.6 และ r = 0.8 เราจะกล่าวถึงความสัมพันธ์ระหว่าง X และ Y ว่าอย่างไร เมื่อเราเปรียบเทียบข้อมูล 2 ชุดนี้เข้าด้วยกัน ($r^2$ = .36 และ $r^2$ = .64)

11.25 บริษัทโฆษณาต้องการทราบว่า จำนวนครั้งที่โฆษณาทางโทรทัศน์ต่อสัปดาห์ และจำนวนขายของสินค้าชนิดหนึ่งต่อสัปดาห์ มีสหสัมพันธ์กันหรือไม่ จึงได้เลือกเมืองตัวอย่าง 9 แห่ง และได้ข้อมูลดังนี้

| เมือง | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| จำนวนโฆษณา (X) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| จำนวนขาย (Y) (100 หน่วย) | 2 | 1 | 3 | 3 | 4 | 5 | 6 | 5 | 7 |

ก) จงหาสัมประสิทธิ์สหสัมพันธ์ (r = .9428)

ข) จงทดสอบ Ho : $\rho = 0$, $\alpha = .05$ (T = 7.48 , ปฏิเสธ Ho)

11.26 งานทดลองเพื่อศึกษาความสัมพันธ์ระหว่างปริมาณน้ำฝน และผลผลิตข้าว ได้ข้อมูลดังนี้

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ปริมาณฝนเป็นนิ้ว (X) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| ผลผลิตเป็นบุชเชล (Y) | 1 | 3 | 2 | 5 | 5 | 4 | 7 | 6 | 9 | 8 |

ก) จงหาสัมประสิทธิ์สหสัมพันธ์

ข) ทดสอบ Ho : $\rho = 0$, $\alpha = 0.05$

11.27 สถิติค่าบำรุงรักษาเครื่องบันทึกและเก็บเงินแบบอัตโนมัติของร้านสรรพสินค้าแห่งหนึ่ง จำนวน 6 เครื่อง มีดังนี้

| | | | | | | |
|---|---|---|---|---|---|---|
| อายุเป็นปี (X) | 2 | 1 | 3 | 2 | 1 | 3 |
| ค่าบำรุงรักษา (Y) | $70 | 40 | 100 | 80 | 30 | 100 |

ก) จงหาสัมประสิทธิ์สหสัมพันธ์ (r = .8702, T = 4.996, ปฏิเสธ Ho)

ข) จงทดสอบว่าอายุ และค่าบำรุงรักษาไม่มีความสัมพันธ์กัน, $\alpha = .05$

11.28 ให้ X คือ GPA ระดับมัธยมปลาย Y = GPA ในมหาวิทยาลัยและกำหนดให้

$\Sigma X = 32.2$ $\Sigma Y = 30.0$ $\Sigma XY = 98.32$

$\Sigma X^2 = 105.74$ $\Sigma Y^2 = 91.90$ $n = 10$

ก) จงหาค่า r (r = .8702)

ข) ทดสอบว่า X และ Y มีความสัมพันธ์กันหรือไม่?, $\alpha = .05$ (T = 4.996, ปฏิเสธ Ho)

11.29 ในการศึกษาความสัมพันธ์ระหว่างคะแนนความถนัด และจำนวนผลผลิตต่อชั่วโมงของคนงานในโรงงานหนึ่งได้ข้อมูลโดยสรุป ดังนี้

$\Sigma X = 550$ $\Sigma Y = 680$ $n = 10$

$\Sigma XY = 45{,}900$ $\Sigma X^2 = 38{,}500$ $\Sigma Y^2 = 56{,}000$

X = คะแนนความถนัด, Y = ผลผลิตต่อชั่วโมง

ก. จงหาสัมประสิทธิ์สหสัมพันธ์ (r = .94725)

ข. จงทดสอบว่า คะแนนความถนัดและผลผลิตต่อชั่วโมงมีความสัมพันธ์กันหรือไม่?

$\alpha = .02$ (T = 8.3599, ปฏิเสธ Ho)

11.30 จำนวนวันหยุดงานในปีที่ผ่านมา (Y) และจำนวนปีที่เข้าทำงาน (X) ของพนักงานที่สุ่มมา 10 คน จากบริษัทขนาดใหญ่แห่งหนึ่ง มีดังนี้

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 2 | 0 | 5 | 6 | 4 | 9 | 2 | 15 | 0 | 7 |
| X | 7 | 8 | 2 | 3 | 5 | 4 | 6 | 10 | 4 | 11 |

ก. จงสร้าง scatter diagram

ข. หาสมการกำลังสองน้อยที่สุด โดยให้ Y เป็นตัวแปรพึ่งพา ($Y = 2.225 + .4625X$)

ค. จงประมาณจำนวนวันหยุดงานของพนักงานผู้หนึ่งซึ่งทำงานมา 4 ปีแล้ว

($Y = 4.075$ วัน)

ง. จงหาสัมประสิทธิ์สหสัมพันธ์ ($r = .30$)

จ. จงทดสอบว่า X และ Y ไม่มีสหสัมพันธ์กันโดยใช้ $\alpha = .01$

(T = 0.89, ยอมรับ Ho, สหสัมพันธ์ไม่มีนัยสำคัญ)

11.31 จำนวนสัปดาห์ที่เข้าทำงาน และจำนวนลูกค้าที่พนักงานสามารถบริการต่อ 1 ชั่วโมงของพนักงานเสริฟอาหาร 5 คน จากภัตตาคารแห่งหนึ่ง มีดังนี้

| จำนวนสัปดาห์ที่เข้าทำงาน (X) | จำนวนลูกค้าที่สามารถให้บริการต่อชั่วโมง (Y) |
|---|---|
| 4 | 10 |
| 7 | 18 |
| 2 | 10 |
| 12 | 28 |
| 5 | 14 |

ก. จงสร้างสมการถดถอย ($\hat{Y} = 4.4138 + 1.931X$)

ข. จงทดสอบ Ho : $\beta = 0$, $\alpha = .05$ (T = 9.165, ปฏิเสธ Ho)

ค. เมื่อ X = 10 จงหาช่วงเชื่อมั่น 95% ของ $\mu_{y/x}$

ง. ค่า $\rho$ ต่างจาก 0 อย่างมีนัยสำคัญที่ $\alpha = .05$ ไหม?

(r = .9826, T = 9.165 , ปฏิเสธ Ho)

11.32 ให้ X คือจำนวนปีที่ทำงาน และ Y คือคะแนนผลงานของพนักงานที่สุ่มมา 10 คน มีดังนี้

| X | 5 | 6 | 5 | 6 | 7 | 8 | 7 | 8 | 9 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 8 | 9 | 6 | 7 | 4 | 7 | 5 | 6 | 3 | 5 |

ก. จงสร้างสมการกำลังสองน้อยที่สุด (Y = 0.75 − 0.75X)

ข. จงหาค่า r (r = −0.61237)

ค. จงทดสอบว่า $\rho$ เป็น 0 หรือไม่ $\alpha = .01$

(T = −2.19, ยอมรับว่า $\rho = 0$)

11.33 จำนวนชั่วโมงที่ศึกษาวิชาสถิติ (X) และคะแนนสอบวิชาสถิติ (Y) ของนักเรียนที่สุ่มมา 6 คน มีดังนี้

| X | 9 | 11 | 14 | 4 | 6 | 6 |
|---|---|---|---|---|---|---|
| Y | 38 | 67 | 49 | 12 | 10 | 24 |

ก. จงสร้างสมการถดถอย (Y = −8.7739 + 5.05289)

ข. จงทดสอบ Ho : $\beta = 0$, Ha : $\beta \neq 0$, $\alpha = .01$

ค. จงหาค่า r (r = .8436)

ง. จงทดสอบ Ho : $\rho = 0$, Ha : P>0, $\alpha = .05$

จ. X อธิบายความผันแปรของ Y ได้เพียงไร? ($r^2 = .71 = 71\%$)

(ข้อ (ข) และ (ง) T = 3.142, ยอมรับ Ho ข้อ (ข) ถ้าใช้ F = 9.87)

11.34 ผลผลิตถั่วเหลือง (Y) และปริมาณน้ำที่ใช้ในระบบชลประทาน (X) สำหรับปลูกถั่วเหลือง 6 แปลง มีดังนี้

| X | 0 | 2 | 4 | 1 | 3 | 5 |
|---|---|---|---|---|---|---|
| Y | 20 | 28 | 32 | 25 | 30 | 31 |

ก. จงสร้างสมการถดถอย (Y = 22.095 + 2.228X)

ข. ถ้าเพิ่มปริมาณน้ำในระบบชลประทาน 1 หน่วย จะมีผลทำให้ผลผลิตเปลี่ยนแปลงไหม?

ค. จงหาค่า $r^2$ และอธิบายความหมาย ($r^2$ = .8577)

ง. จงทดสอบว่าตัวแปรคู่นี้ไม่มีความสัมพันธ์กัน $\alpha$ = .01 (F = 24.11, ปฏิเสธ Ho)

จ. เมื่อ X = 5 จงหาช่วงเชื่อมั่นของ $\mu_{y/x}$ และ $Y'_a$, $\alpha$ = .05 ($29.4 < \mu_{y/x} < 37.05$ ; $26.7 < Y_a < 39.7$)

11.35 ให้ X = ปริมาณปุ๋ยไนโตรเจน Y = ผลผลิตข้าว

n = 20 $\Sigma Y$ = 260

$\Sigma X$ = 230 $\Sigma X^2$ = 3144

$\Sigma XY$ = 3490 $\Sigma Y^2$ = 3904

ก. จงสร้างสมการถดถอย (Y = 1.4769 + 1.00204X)

ข. จงทดสอบ Ho : $\beta$ = 0, Ha : $\beta > 0$, $\alpha$ = .01 (T = 19.8 ปฏิเสธ Ho)

ค. จงทดสอบ Ho : $\rho$ = 0 , Ha : $\rho \neq 0$, $\alpha$ = .10 (T = 19.8, ปฏิเสธ Ho)

ง. จงหาช่วงเชื่อมั่น 98% ของ $\beta$ ($.873 < \beta < 1.131$)

11.36 ในทางธุรกิจเราทราบว่า ถ้าซื้อสินค้าจำนวนมาก ราคาต่อหน่วยจะลดลง ดังข้อมูลที่เก็บมา ซึ่งมี X = ปริมาณสินค้าเป็นหน่วยที่ซื้อ และ Y = ราคาต่อหน่วย

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Y | 10 | 8 | 7 | 6 | 4 |

ก) จงหาจุดตัดที่แกน Y และความลาดชันของเส้นถดถอย (a = 11.2, b = −1.4)

ข) จงทดสอบ Ho : $\beta$ = 0, Ha : $\beta \neq 0$, $\alpha$ = .05 (F = 147, ปฏิเสธ Ho)

ค) จงหาค่า r (r = −.9899)

ง) มีหลักฐานเพียงพอที่จะปฏิเสธ Ho ว่า ไม่มีความสัมพันธ์กันระหว่าง X และ Y หรือไม่? $\alpha$ = .05 (คำตอบในข้อ (ข))

11.37 ให้ X คือ คะแนนปีสุดท้าย (GPA) ของผู้จบมหาวิทยาลัย และ Y คือ เงินเดือนขั้นต้นของบัณฑิต 30 คน และมี computor printout ดังนี้

MEANS AND STANDARD DEVIATIONS

X : 2.83833 + − 0.44841 Y : 10740.66667 + − 1967.65248

CORRELATIONS

PEARSON : 0.827368 SPEARMAN : 0.933515 **KENDALL** : 0.799219

LINEAR REGRESSION OF Y ON X

SLOPE : 3630.56128 + – 465.76874

INTERCEPT : 435.92357 + – 1337.85967

S(Y/X) : 1124.71499

ANALYSIS OF VARIANCE : F (1, 28) = 60.75847

ก. จงสร้างสมการถดถอย (Y = 435.92357 + 36.30.56128X)

ข. จงหา 95% ช่วงเชื่อมั่นของ $\beta$

ค. จะปฏิเสธ Ho : $\beta$ = 4000 ที่ $\alpha$ = .05 ได้ไหม? (T = −0.7932, ยอมรับ Ho)

ง. จะปฏิเสธ Ho : $\mu_{y/x}$ = 11,500 เมื่อ Xo = 2.75 ได้ไหม? $\alpha$ = .05

(T = 5.157, ปฏิเสธ Ho)

จ. จงหา 95% ช่วงเชื่อมั่นของ $\mu_{y/X_O}$ เมื่อ Xo=2.75 (99991.065<$\mu_{y/xo}$<10,848.867)

ฉ. GPA อธิบายความผันแปรของเงินเดือนเริ่มต้นได้กี่เปอร์เซ็นต์? (68.45%)

**หมายเหตุ** ค่า r ที่เราศึกษาในบทที่ 11 คือ ค่าสัมประสิทธิ์สหสัมพันธ์แบบเบียร์สัน (Pearson carrelation coeffiecient) ส่วนแบบ Spearman และแบบ Kendall เป็นแบบ Nonparametric ซึ่งจะได้ศึกษาต่อไป

11.38 กำหนดให้ a = 7.73750, b = 24.77500, MSE = 27.99583, n = 8, $\bar{X}$ = 5.5, $\Sigma(X - \bar{X})^2$ = 40.0 และ $S^2_{\hat{y}}$ สำหรับค่าต่างของ $\hat{Y}$ ดังนี้

| $X_o$ | : | 5 | 6 | 7 |
|---|---|---|---|---|
| $S^2_{\hat{y}}$ | : | 3.674 | 3.674 | 5.074 |

ก. เหตุใด $S^2_{\hat{y}}$ เมื่อ $X_O$ = 7 จึงต่างหาก 2 อันแรก

ข. จงแสดงว่า เมื่อ $X_o$ = 6, $S^2_{\hat{y}}$ = 3.674

11.39 เหตุใดเมื่อทดสอบ Ho : $\beta = 0$, Ha : $\beta \neq 0$ โดย F-test จึงเป็นการทดสอบแบบด้านเดียว ทั้ง ๆ ที่สมมติฐานรอง หมายถึง $\beta > 0$ หรือ $\beta < 0$

11.40 โรงงานผลิตภัณฑ์นมและเนยแข็งต้องการทราบความสัมพันธ์ของต้นทุนการผลิต และผลผลิตเนยแข็ง เขามีข้อมูลดังนี้

| ผลผลิต | Y = ต้นทุนรวม | X = จำนวนผลผลิต (ปอนด์) |
|---|---|---|
| 1 | 300 | 200 |
| 2 | 500 | 1,000 |
| 3 | 450 | 600 |
| 4 | 280 | 350 |
| 5 | 470 | 800 |
| 6 | 400 | 700 |
| 7 | 390 | 500 |
| 8 | 200 | 150 |
| 9 | 450 | 900 |
| 10 | 360 | 400 |

$\Sigma X = 5,600; \Sigma Y = 3,800; \Sigma X^2 = 3,905,000; \Sigma XY = 2,358,000;$
$\Sigma Y^2 = 1,526,000$

จงหาสมการกำลังสองน้อยที่สุดเพื่อช่วยโรงงานประมาณต้นทุนการผลิตโดยกำหนดน้ำหนักของผลผลิต ($Y = 212.50977 + 0.2990897X$)

11.41 จากข้อ 11.40 จงหา $S_{y/x}$ (หรือ $S_E$ = standard error of estimate) และสัมประสิทธิ์การตัดสินใจ และจงอธิบายความหมายของค่าทั้งสอง ($S_{y/x} = 40.634604, r^2 = .8389$)

11.42 จากข้อ 11.40 ภายหลังจากโรงงานได้สมการประมาณค่าแล้ว โรงงานได้พบงานวิจัยชิ้นหนึ่ง ซึ่งอ้างว่าเมื่อเพิ่มจำนวนผลิต 1 ปอนด์ จะทำให้ต้นทุนเพิ่มขึ้น $0.24 (incremental cost) เมื่อเปรียบเทียบกับข้อมูลที่ได้ในข้อ 11.40 จะมีความขัดแย้งกันว่า incremental cost ต่างจาก $0.24 ต่อปอนด์ที่ระดับนัยสำคัญเท่าใด (จงหา p-value)

11.43 จงสร้างช่วงเชื่อมั่น 80% ของต้นทุนการผลิตเมื่อกำหนดน้ำหนักให้ (ต่อจากข้อ 11.40)

11.44 จากข้อ 11.40 จงทดสอบ Ho : $\beta = 0$ (T = 6.45; **ปฏิเสธ** Ho)

11.45 จากข้อ 11.40 จงสร้างช่วงเชื่อมั่น 90% ของต้นทุนถัวเฉลี่ยสำหรับการผลิต 500 ปอนด์ ($337.6 < \mu_{y/x_o} < 386.5$)

11.46 โรงงานต้องการประมาณต้นทุนการผลิตเนย 500 ปอนด์ เพื่อจะได้ยื่นประมูลกับโรงแรมแห่งหนึ่ง และต้องการทราบช่วงเชื่อมั่น 80% ของค่าประมาณ จงหาช่วงเชื่อมั่นดังกล่าว ($302.39 < Y_a < 421.72$)

---