ใบความรู้ที่ 12
เรื่อง การวิเคราะห์ความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูล

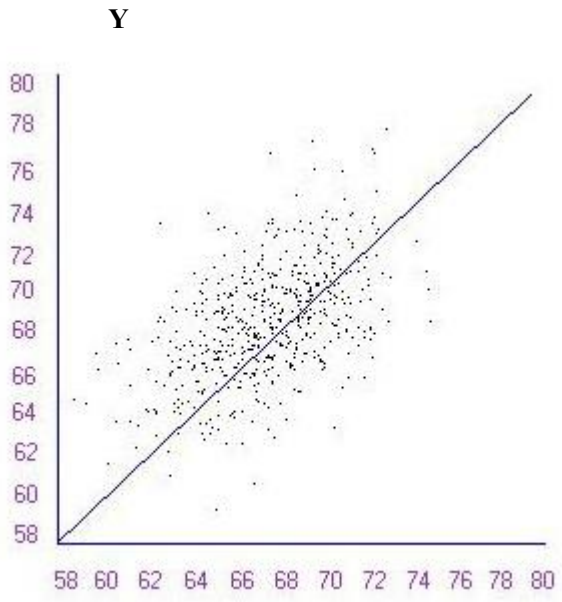
ในบทที่ 1 และในบทที่ 2 เป็นการศึกษาข้อมูลเชิงปริมาณของตัวแปรเดียวหรือของแต่ละตัวแปรในข้อมูลแต่ละชุด ส่วนในบทที่ 3 นี้ เป็นการศึกษาข้อมูลเชิงปริมาณของตัวแปรทีละสองตัวพร้อมกัน โดยมุ่งตรวจสอบความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูลของสองตัวแปรนั้น ๆ เพื่อประโยชน์ในการใช้ตัวแปรตัวหนึ่งไปพยากรณ์ตัวแปรอีกตัวหนึ่งภายใต้สมการที่แสดงความสัมพันธ์เชิงฟังก์ชันที่คำนวณได้ โดยทั่วไปสมการที่ใช้แสดงความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูลอาจประกอบด้วยตัวแปรหลายตัว แต่ในระดับเบื้องต้นนี้เป็นการศึกษาในกรณีพื้นฐานของสองตัวแปรเท่านั้น

สมการที่ใช้แสดงความสัมพันธ์เชิงฟังก์ชันในบทที่ 3 นี้ มีรูปแบบของความสัมพันธ์ของกราฟที่เป็นแบบเส้นตรงและไม่เป็นเส้นตรง ดังนั้น การเลือกใช้สมการแบบใดจึงจำเป็นต้องตรวจสอบโดยการลงจุดแผนภาพการกระจาย (scatterplots) ระหว่างข้อมูลของ 2 ตัวแปรก่อน เพื่อตรวจดูรูปแบบที่เหมาะสม แล้วจึงวิเคราะห์ข้อมูลในลำดับต่อไป

การวิเคราะห์ความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูล

ในการวิเคราะห์ข้อมูลบ่อยครั้งมีข้อมูลเชิงปริมาณที่ประกอบด้วยตัวแปรตั้งแต่สองตัวขึ้นไป และตัวแปรเหล่านั้นมีความเกี่ยวข้องกันอยู่ เช่น รายได้และรายจ่ายของครอบครัว ส่วนสูงและน้ำหนักของเด็กแรกเกิด ความเกี่ยวข้องกันของตัวแปรจะมีลักษณะที่ค่าของตัวแปรตัวหนึ่งขึ้นอยู่กับอีกตัวแปรหนึ่ง เช่น รายจ่ายจะขึ้นอยู่กับรายได้ ส่วนสูงขึ้นอยู่กับน้ำหนัก กรณีเช่นนี้จะเรียกตัวแปรที่แสดงรายได้หรือน้ำหนักว่า **ตัวแปรอิสระ (independent variables)** เรียกตัวแปรที่แสดงรายจ่ายหรือส่วนสูงว่า **ตัวแปรตาม (dependent variables)**

Karl Pearson เป็นผู้หนึ่งที่ศึกษาเรื่องความคล้ายคลึงกันของสมาชิกในครอบครัว ในปี ค.ศ. 1903 เขาวัดความสูงของบิดาจำนวน 1,078 คน และความสูงของบุตรชายคนหนึ่งที่เติบโตเต็มที่ของบุคคลเหล่านี้ นำความสูงของบิดาและบุตรจำนวน 1,078 คู่นี้ มาสร้างแผนภาพการกระจายดังภาพที่ 1 โดยกำหนดแกนนอนหรือแกน x แทนความสูงของบิดา แกนตั้งหรือแกน y แทนความสูงของบุตรชาย และแต่ละจุดแทนคู่บิดาและบุตรชายหนึ่งคู่



ภาพที่ 1 แผนภาพการกระจายของความสูงของบิดาและบุตรชาย 1,078 คู่

จากภาพ 1 แสดงให้เห็นความเกี่ยวข้องกันระหว่างสองตัวแปรคือความสูงของบิดาและความสูงของบุตรชาย โดยจะเห็นกลุ่มของจุดที่เอียงสูงขึ้นทางด้านขวามือ กล่าวคือ ค่า y ของจุดส่วนใหญ่จะเพิ่มขึ้นตามค่า x ที่เพิ่มขึ้น หมายความว่า บิดาที่สูงมักจะมีบุตรชายที่สูงด้วย นักสถิติกล่าวถึงลักษณะเช่นนี้ว่า ความสูงของบิดาและบุตรชายมีสหสัมพันธ์กันในทางบวก

คำว่า สหสัมพันธ์ (correlation) แยกเป็นคำ 2 คำ คือ **สห** ซึ่งหมายถึง ร่วมกันหรือด้วยกัน และ **ความสัมพันธ์** หมายถึง ความเกี่ยวข้องกัน เมื่อเหตุการณ์ 2 เหตุการณ์ที่โดยปกติมักเกิดขึ้นพร้อมกัน จะบอกว่าสองเหตุการณ์นั้นมีสหสัมพันธ์กัน เช่น คนผมสีดำและตาสีน้ำตาล คนผมสีทองและตาสีฟ้า นอกจากนี้เมื่อมีการเปลี่ยนแปลงในเหตุการณ์หนึ่ง ก็มักเกิดการเปลี่ยนแปลงในอีกเหตุการณ์หนึ่งควบคู่กัน เช่น เมื่อเด็กสูงขึ้น เขาน่าจะมีน้ำหนักเพิ่มขึ้น

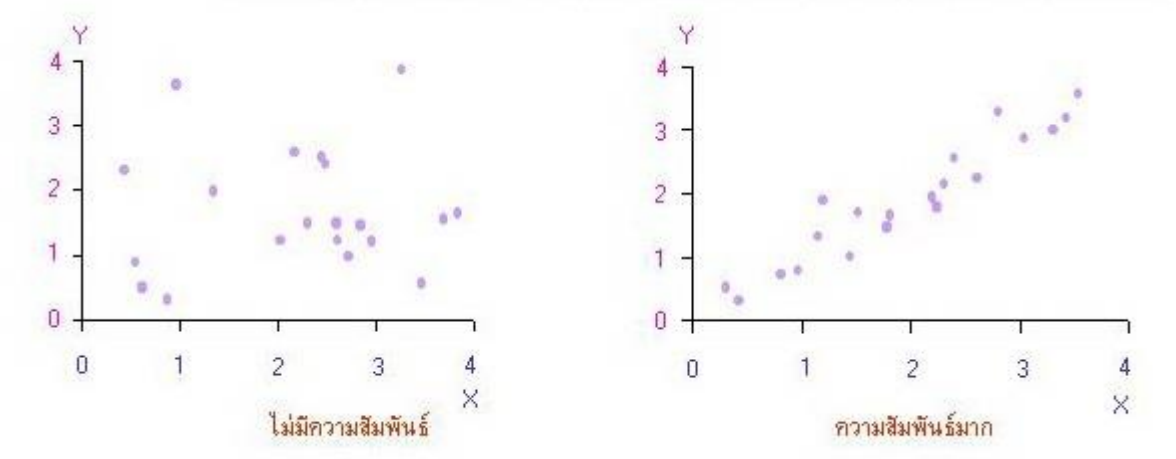
สหสัมพันธ์มี 2 แบบ คือ สหสัมพันธ์ทางบวกและสหสัมพันธ์ทางลบ สหสัมพันธ์ทางบวกหมายถึง เมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้น อีกตัวแปรมีค่าเพิ่มขึ้นตาม ส่วนสหสัมพันธ์ทางลบ หมายถึง เมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้น อีกตัวแปรจะมีค่าลดลง

เมื่อทราบจากแผนภาพการกระจายว่าตัวแปรมีสหสัมพันธ์กัน สิ่งที่ควรทราบเพิ่มเติมคือ ความเกี่ยวข้องสัมพันธ์นั้นมีมากหรือน้อยเพียงใด ในเรื่องนี้แผนภาพการกระจายจะสามารถบอกได้ในระดับหนึ่ง เมื่อกล่าวถึงบิดาที่สูง 72 นิ้ว อาจคาดได้ว่าบุตรชายจะสูง 72 นิ้วด้วย ในทำนองเดียวกัน ถ้าบิดาสูง 68 นิ้ว คาดว่าบุตรชายควรสูง 68 นิ้ว หรือถ้าบิดาสูง 70 นิ้ว บุตรชายก็น่าจะสูง 70 นิ้ว นั่นคือ หากนำความสูงของบิดาและบุตรชายคู่ต่าง ๆ เหล่านี้มาลงจุดในแผนภาพ จุดจะตกบนเส้นตรงที่ทำมุม 45^{o} กับแกนนอน เส้นตรงนี้เป็นเส้นที่แสดงว่าความสูงของบุตรชายเท่ากับความสูงของบิดา โดยมีสมการเป็น $y = x$ ดังภาพที่ 1

ฉะนั้นถ้าคิดว่าความสูงของบุตรชายควรใกล้เคียงกับความสูงของบิดา หมายความว่า จุดต่าง ๆ บนแผนภาพการกระจายควรตกใกล้กับเส้นตรงเส้นนี้ ซึ่งจากภาพที่ 1 จะเห็นครอบครัวส่วนใหญ่มีจุดตกกระจายรอบ ๆ เส้น บ้างก็ห่างจากเส้นตรงมาก บ้างก็อยู่ใกล้เคียง แสดงว่าความสูงของบุตรชายต่างจากความสูงของบิดาไม่มากก็น้อย

การกระจายของจุดในแผนภาพการกระจายแสดงถึงความมากหรือน้อยของความสัมพันธ์ระหว่างความสูงของบิดาและบุตรชาย การทราบความสูงของบิดาช่วยให้คาดเดาความสูงของบุตรชายได้ เพราะความสูงของบิดาและบุตรชายมีความสัมพันธ์กัน

แต่การคาดคะเนก็ไม่ถูกต้องแน่นอน ยังมีความผิดพลาดเกิดขึ้นได้ เพราะบุตรชายที่มีบิดาสูงเท่ากันหลายคนก็มีความสูงแตกต่างกัน ลองพิจารณาบิดาที่สูงประมาณ 72 นิ้ว ในภาพที่ 1 จุดต่าง ๆ ที่มีค่า x ใกล้ ๆ 72 นิ้วล้วนเป็นจุดจากคู่บิดาและบุตรชายที่มีบิดาสูง 72 นิ้ว จะเห็นว่าความสูงของบุตรชายเหล่านี้ (ค่า y) มีการกระจายหรือความผันแปรอยู่มาก นั่นคือ การทำนายความสูงของบุตรชายมีความคลาดเคลื่อนได้พอสมควร ถึงแม้ว่าจะทราบความสูงของบิดาของเขา อันเนื่องมาจากความสัมพันธ์ระหว่างตัวแปร ทั้งสองยังไม่สมบูรณ์ ดังนั้นจะสามารถสรุปความสัมพันธ์ของตัวแปร x และ y ออกมาเป็นตัวเลขให้เห็นว่ามีระดับมากหรือน้อยได้อย่างไร



ภาพที่ 2 การกระจายของข้อมูล 2 ชุด ที่มีค่ากลางและการกระจายเหมือนกันแต่ระดับความสัมพันธ์ต่างกัน

ค่าเฉลี่ยของ x และ y รวมทั้งส่วนเบี่ยงเบนมาตรฐานของ x และ y ไม่อาจอธิบายเกี่ยวกับความสัมพันธ์ระหว่าง x และ y ได้ ค่าเฉลี่ยของ x และ y จะแสดงให้ทราบว่าจุดศูนย์กลางของกลุ่มข้อมูลอยู่ที่ใด และส่วนเบี่ยงเบนมาตรฐานของ x และ y จะอธิบายเรื่องการกระจายของจุดบนแต่ละแกน จากด้านหนึ่งของกลุ่มไปยังอีกด้านหนึ่ง พิจารณาแผนภาพการกระจายของข้อมูล 2 ชุดในภาพที่ 2 เห็นได้ว่าทั้งสองชุดต่างมีจุดศูนย์กลางและการกระจายด้านแกนนอนและแกนตั้งเหมือนกัน แต่ในชุดแรก จุดกระจัดกระจายไม่เกาะกลุ่มกัน ส่วนในชุดที่สองจุดเกาะกลุ่มแนบแน่นเป็นแนวเส้นตรงมาก หรือสองตัวแปรมีความสัมพันธ์เชิงเส้นตรงสูงมาก นั่นคือ ระดับความสัมพันธ์ในแผนภาพทั้งสองนี้ต่างกัน การจะวัดระดับความสัมพันธ์ จึงต้องใช้ค่าทางสถิติอีกค่าหนึ่งที่เรียกว่า สัมประสิทธิ์สหสัมพันธ์ (correlation coefficient)

ทั้งนี้ จุดประสงค์สำคัญของการสร้างความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูลสองชุด ก็เพื่อใช้สมการความสัมพันธ์เชิงฟังก์ชันของ x และ y ในการพยากรณ์ค่าของตัวแปรตาม เมื่อทราบค่าของตัวแปรอิสระ เช่น ถ้าทราบความสูงของบิดา (x) ก็สามารถพยากรณ์ความสูงของนักเรียน (y) ได้โดยแทนค่า x ลงในสมการแล้วคำนวณค่าของ y นั่นเอง

ในการสร้างความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูลเชิงปริมาณนี้ ข้อมูลเชิงปริมาณที่นำมาสร้างความสัมพันธ์จะต้องประกอบด้วยค่าจากการสังเกตเป็นจำนวนมากพอสมควร เช่น ตั้งแต่ 10 ค่าขึ้นไป เพราะถ้าค่าจากการสังเกตมีจำนวนน้อยแล้วความสัมพันธ์เชิงฟังก์ชันระหว่างข้อมูลเชิงปริมาณของตัวแปรสองตัวที่สร้างขึ้น อาจจะไม่สามารถแทนความสัมพันธ์ที่ควรจะเกิดขึ้นจริง ๆ ระหว่างตัวแปรทั้งสอง จะเป็นผลทำให้การพยากรณ์ค่าของตัวแปรตามที่ต้องการทราบอาจคลาดเคลื่อนไปจากที่ควรจะเป็นจริงมาก

โดยทั่ว ๆ ไป ความสัมพันธ์เชิงฟังก์ชันของข้อมูลที่ประกอบด้วยตัวแปรสองตัวแปรอาจแบ่งออกเป็นสองชนิดใหญ่ ๆ คือ

(1) ความสัมพันธ์เชิงฟังก์ชันที่กราฟเป็นเส้นตรง มีสมการทั่วไปของความสัมพันธ์เชิงฟังก์ชันเป็น

$Y = a + bX$

เมื่อ Y เป็นตัวแปรตาม และ X เป็นตัวแปรอิสระ

b เป็นความชันของเส้นตรงหรือค่าของ Y ที่เปลี่ยนไปเมื่อ X เปลี่ยนไปหนึ่งหน่วย

a เป็นระยะตัดแกน Y และเป็นค่าคงตัวที่ต้องการหา

(2) ความสัมพันธ์เชิงฟังก์ชันที่กราฟไม่เป็นเส้นตรง ในที่นี้จะกล่าวเฉพาะความสัมพันธ์ที่มีกราฟเป็นรูปพาราโบลา และรูปเอกซ์โพเนนเชียล ซึ่งมีรูปสมการของความสัมพันธ์เชิงฟังก์ชัน ดังนี้

$Y = a + bX + cX^2$ (รูปพาราโบลา)

$Y = ab^X$ (รูปเอกซ์โพเนนเชียล)

เมื่อ Y เป็นตัวแปรตาม

X เป็นตัวแปรอิสระ

a, b และ c เป็นค่าคงตัวที่จะต้องหา
