บทที่ 2

ความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง

1. ความรู้พื้นฐาน

การวิจัยเพื่อศึกษาสถิติวิเคราะห์ปริมาณน้ำตาลในเนื้อผลสับปะรด ผู้วิจัยได้ศึกษาเอกสารแนวคิดทางวิชาการเพื่อใช้เป็นแนวทางในการกำหนดกรอบการศึกษาดังนี้

1. วิธี เอนไออาร์ สเปกโตรสโกปี

2. สถิติวิเคราะห์เพื่อสร้างสมการ calibration

3. สถิติวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างปริมาณน้ำตาลจากสมการcalibration กับปริมาณน้ำตาลที่วัดจากกรรมวิธีทางเคมี

ซึ่งจะขอกล่าวถึงเป็นเรื่องๆ ไป

1.) วิธี เอนไออาร์ สเปกโตรสโกปี

เอนไออาร์ สเปกโตรสโกปี เป็นวิธีการวัดการดูดกลืนแสงในช่วงคลื่น near infared (NIR) ที่จะทำให้โมเลกุลของสารเกิดการสั่นสะเทือน ข้อมูลที่ได้จากการดูดกลืนแสง NIR ที่ความยาวคลื่นต่างๆ สามารถบอกลักษณะเฉพาะของสารและในขณะเดียวกันจะบอกถึงปริมาณของสารนั้น เทคนิค NIR นี้สามารถนำมาประยุกต์ในการตรวจสอบคุณภาพของอาหาร โดยไม่ก่อให้เกิดผลเสียแก่อาหารได้

1.1 ลักษณะสำคัญของ เอนไออาร์ สเปกโตรสโกปี

1.1.1 ไม่ใช้สารเคมีจำนวนมากเหมือนอย่างที่ต้องใช้ในการวิเคราะห์ทางเคมีทั่วไป ซึ่งนอกจากจะมีประโยชน์ที่ประหยัดค่าใช้จ่ายแล้ววิธีนี้ยังไม่ก่อให้เกิดมลพิษแก่ห้องปฏิบัติการ

1.1.2 ทำให้วิเคราะห์ได้อย่างสะดวกรวดเร็วด้วยการเตรียมตัวอย่างที่ใช้ทดสอบอย่างง่ายดาย

1.1.3 ไม่ต้องการช่างเทคนิคที่ต้องผ่านการอบรมพิเศษเรื่องขั้นตอนการวิเคราะห์ถ้าระบบวิเคราะห์โดยสมบูรณ์ใช้ได้โดยง่าย

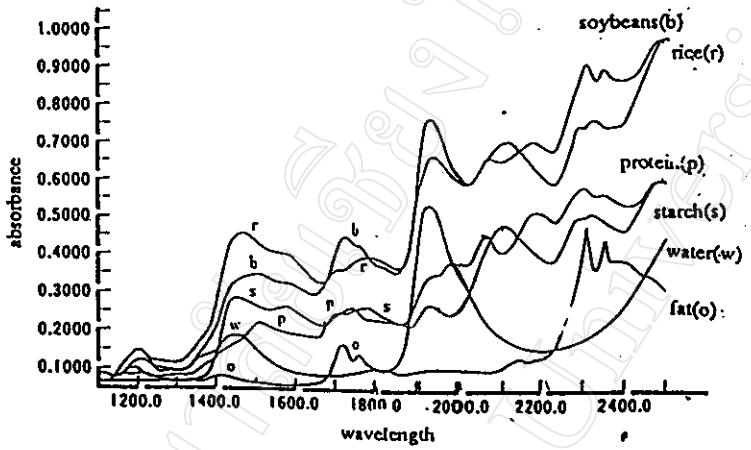
1.1.4 ตัวอย่างเดียวสามารถนำมาทดสอบได้หลายครั้ง

1.1.5 สามารถทราบข้อมูลได้มากกว่า 1 ชนิดจากการวัดเพียงครั้งเดียว

1.1.6 วิเคราะห์ได้รวดเร็วในขณะนั้นและยังเหมาะสมสำหรับการควบคุมคุณภาพ (QC) ที่ใช้ในโรงงาน

1.1.7 สามารถตรวจสอบผลิตผลได้ทุกชนิดโดยการดัดแปลงการวัดเป็น แบบการติดต่อได้ทุกที่อย่างทั่วถึงตลอดเวลา(on-line)

1.1.8 สามารถใช้ในสถานที่ที่มีการสั่นสะเทือนหรือเคลื่อนไหวได้ เช่น บนเรือ เพราะไม่ต้องอาศัยการชั่งน้ำหนัก



ภาพที่ 1 สเปกตรัมการดูดกลืน NIR ของข้าว,ถั่วเหลือง และส่วนประกอบ

ภาพที่ 1 แสดงถึงสเปกตรัมของ NIR (NIR spectra) ของถั่วเหลืองและข้าว รวมทั้งส่วนประกอบหลัก เช่น น้ำ, โปรตีน, ไขมัน และแป้ง ยิ่งแสงถูกดูดซับมากเท่าใดก็จะยิ่งกลายเป็นค่าการดูดซับมากเท่านั้น แถบการดูดซับของแต่ละส่วนประกอบขึ้นอยู่กับลักษณะเฉพาะกลุ่ม แถบดูดซับที่ 1935 nm ที่สังเกตพบทั้งในข้าวและถั่วเหลืองเกิดจากน้ำ แถบที่ 2100 nm ที่พบในข้าวเกิดจากแป้ง ซึ่งไม่ได้สังเกตพบอย่างชัดเจนในถั่วเหลืองเนื่องจากมีแป้งอยู่น้อย แถบการดูดซับของโปรตีนที่ 2180 nm และแถบของไขมันที่ 2305 และ 2345 nm พบได้ชัดเจนในถั่วเหลืองซึ่งมีโปรตีนและไขมันมาก

จะเห็นได้ว่า เอนไอาร์ สเปกโตรสโกปี สามารถนำมาใช้ในการประเมินคุณภาพอาหารที่ดีที่สุดในปัจจุบัน ซึ่งสามารถวัดคุณภาพอาหารได้อย่างรวดเร็วและไม่ทำให้อาหารเกิดความเสียหาย นอกจากนี้สำหรับผู้บริโภค การวัดคุณภาพอาหารด้วยวิธีนี้ จะทำให้สามารถเลือกรสชาติได้ตามที่ชอบและเลือกซื้อผลผลิตที่มีคุณภาพเหมือนกันได้ ซึ่งจะทำให้ผู้ผลิตได้รับผลตอบแทนมากน้อย ตามแต่คุณภาพของผลิตผลที่พวกเขาจำหน่าย ส่งผลให้ผู้ผลิตต้องตระหนักในคุณภาพของผลผลิตของตน และเกิดการปรับปรุงเทคนิคในการเพาะปลูกให้ดียิ่งขึ้น

1.2 หลักการของ เอนไออาร์ สเปกโตรสโกปี (บริษัทจาร์พา เท็คเซ็นเตอร์ จำกัด, 2540)

โดยทั่วไปแล้ว Near Infared เป็นรังสีที่มีช่วงความยาวคลื่นอยู่ระหว่าง 800 นาโนเมตร ถึง 2500 นาโนเมตร ซึ่งการดูดกลืนคลื่นแม่เหล็กไฟฟ้าโดยสาร(หรือการดูดกลืนโฟตอน) ในช่วงคลื่นนี้จะทำให้โมเลกุลของสารสั่นด้วยความถี่สูงขึ้น นั่นคือโมเลกุลของวัตถุถูกกระตุ้นให้เปลี่ยนระดับพลังงานจาก "สถานะพื้น (ground state)" ไปสู่ "สถานะกระตุ้น(excited state)" ในทางฟิสิกส์ทฤษฎีควอนตัมพบว่าโมเลกุลจะมีระดับพลังงานของการสั่น(E) เป็น

$$E_n = (n + \tfrac{1}{2})h\nu$$

เมื่อ h คือ ค่าคงที่ของ Planck (เท่ากับ $6.626*10^{-34}$ จูล-วินาที) (Krane, S.K. ,1983)

ν คือ ความถี่ของการสั่น

n คือ เลขควอนตัมของการสั่น (n=0,1,2,...)

ในช่วง NIR โมเลกุลจะถูกกระตุ้นจากระดับ n=0 (สถานะพื้น:สถานะที่มีความคงที่ที่สุด) ไปสู่ระดับ n=2 (First overtone, 1,350 nm - 8,000 nm) หรือระดับ n=3 (second overtone, 900 - 1,200 nm) ซึ่งการส่งผ่านไปยังระดับพลังงาน n=1 จะอยู่ในช่วงรังสีใต้แดง(infared ,IR:2,700 - 8,000 nm)

ปริมาณการดูดกลืนคลื่นแม่เหล็กไฟฟ้า ที่ความยาวคลื่นใดๆ (λ) ของสารตัวอย่าง สามารถวัดออกมาในรูปของ "ค่าการส่งผ่าน(transmittance:T) โดยที่

$$T = I/I_0$$

เมื่อ

I_0 คือ ความเข้มรังสี ก่อนผ่านสารตัวอย่าง

I คือ ความเข้มรังสี หลังผ่านสารตัวอย่าง

ค่าของ T อยู่ระหว่าง 0 (ดูดกลืนสมบูรณ์) ถึง 1 (ไม่มีการดูดกลืน) สำหรับวัตถุตัวอย่างต่างๆ "ค่าการดูดกลืน(Absorbance:A)" นิยามโดย

$$A = -\log(T) = \log(1/T)$$

ค่าการดูดกลืนนี้มีความสัมพันธ์เชิงเส้นกับความเข้มข้นของสารประกอบในตัวอย่างที่ดูดกลืนแสง และยังสามารถใช้กฎของ Beer-Lambert คือ

$$A = \varepsilon lC$$

เมื่อ

l คือ ความหนาของวัตถุตัวอย่าง

C คือ ความเข้มข้นของสารตัวอย่างชนิดที่ดูดกลืนรังสี (ปกติใช้หน่วย mole,M)

ε คือ ค่าสัมประสิทธิ์การดูดกลืนของสารตัวอย่าง

ในช่วงคลื่น IR (infrared)โดยปกติแล้วสารประกอบจะมีคุณสมบัติการดูดกลืนสูงมาก (คือ สามารถดูดกลืนรังสี IRได้ดี) และความหนาของสารตัวอย่างค่อนข้างจะมีค่าน้อยมาก(น้อยกว่า 1 มิลลิเมตร) นอกจากนี้ ค่าการส่งผ่านจะเป็นศูนย์ (ค่าการดูดกลืน A=∞) ทำให้ไม่สามารถคำนวณหาความเข้มข้นได้ แต่สำหรับในคลื่น NIR ค่าการดูดกลืนจะมีค่าต่ำมาก และมีระยะเคลื่อนผ่าน (pathlength) ที่ยาว (เซนติเมตร) และสามารถคำนวณความเข้มข้นของสารตัวอย่างออกมาได้

อย่างไรก็ตาม ถ้าค่าการดูดกลืนมีค่ามาก(หรือสารตัวอย่างหนา) แล้วจะไม่มีแสงทะลุผ่านสารตัวอย่างนั้น ความเข้มของแสงที่ตกกระทบพื้นผิวของสารตัวอย่างจะสะท้อนกลับหมด ซึ่งเรียกว่า "ค่าการสะท้อนแบบแพร่(Diffuse Reflectance; R)" ของสารตัวอย่าง สามารถวัดได้ในสเปกตรัมช่วง NIR (ที่ความยาวคลื่น λ) โดยนิยาม ได้เป็น

$$R = I_S / I_R$$

เมื่อ

I_S คือ ความเข้มของรังสีที่สะท้อนจากสารตัวอย่าง

I_R คือ ความเข้มของรังสีที่สะท้อนจากวัตถุอ้างอิง(ให้พื้นผิวแบบสะท้อน 100 %)

กฎของ Beer-Lambert สามารถใช้ได้กับค่า $\log(1/R)$ ซึ่งเป็นค่าความดูดกลืนของสารตัวอย่างได้เช่นเดียวกัน

ถ้าต้องการหาปริมาณส่วนประกอบเฉพาะในสารตัวอย่างชนิดหนึ่ง (ตัวอย่างเช่น ความชื้น, โปรตีน, ไขมัน ฯลฯ) จะเลือกความยาวคลื่น λ^* ที่ค่าการดูดกลืนแสง เป็นผลอันเนื่องมาจากสารประกอบของสารดังกล่าวเท่านั้น (เช่น น้ำ, drug molecule) ดังนั้น

$$A(\lambda^*) = \varepsilon(\lambda^*)lC$$

เมื่อ

$A(\lambda^*)$ คือ ค่าการดูดกลืนที่ความยาวคลื่นใดๆ

l คือ ความหนาของวัตถุตัวอย่าง

$\varepsilon(\lambda^*)$ คือ ค่าสัมประสิทธิ์การดูดกลืนของสารตัวอย่างที่ความยาวคลื่นใดๆ

ค่าของ $\varepsilon(\lambda^*)$ สามารถหาได้จากการวัดค่าความดูดกลืนของสารอ้างอิงมาตรฐานที่ทราบค่าความหนาแน่นของสารตัวอย่าง ดังนั้นจะสามารถหาค่า C ได้

อย่างไรก็ตามในย่าน NIR ชนิดโมเลกุลส่วนใหญ่ดูดกลืนแสงที่ความยาวคลื่นเดียวกัน แต่มีค่าการดูดกลืนแตกต่างกัน ตรงนี้จึงเป็นสาเหตุให้ค่าการดูดกลืนเป็นผลเนื่องมาจากการสั่นของ functional group ต่างๆ เช่น –OH, -NH, -CH, และสารที่มีส่วนประกอบเป็นสารอินทรีย์ ด้วยเหตุนี้การวัดค่าการดูดกลืนจึงรวมไปถึงองค์ประกอบทุกอย่างในสารตัวอย่าง โดยใช้วิธีวัดค่าหลายตัวแปร (multivariate calibration) ในการวิเคราะห์ข้อมูล

2.) สถิติวิเคราะห์เพื่อสร้างสมการ calibration

สถิติวิเคราะห์เชิงปริมาณที่เกี่ยวข้องกับ เอนไอาร์ สเปกโตรสโกปี มีหลายวิธี สำหรับในการวิจัยครั้งนี้ได้ใช้วิธี Principal Component Regression (PCR)

Principal Component Regression (PCR) (Galacitc Industries Corporation, 1985)

วิธีการนี้เป็นการรวมเอาวิธี Principal Component Analysis และ Inverse Least Square Regression ไว้ด้วยกันเพื่อใช้แก้สมการ calibration กระบวนการวิเคราะห์ PCR จึงแบ่งเป็น 2 ขั้นตอน

1. การวิเคราะห์องค์ประกอบหลัก (PCA)
2. Inverse Least Square Regression

2.1 การวิเคราะห์องค์ประกอบหลัก

PCA เป็นเทคนิคการสกัดปัจจัยวิธีหนึ่งในหลายๆวิธีของการวิเคราะห์ปัจจัย โดยมีวัตถุประสงค์เพื่อลดข้อมูล (ตัวแปร) ให้น้อยลงโดยอาศัยหลักความสัมพันธ์เชิงเส้นระหว่างตัวแปร (linear combination of the observed data) ที่ใช้เป็นข้อมูลแต่ไม่มีการสมมติเกี่ยวกับความสัมพันธ์เชิงสาเหตุและผลระหว่างปัจจัยและตัวแปร หรือ มีเพื่อลดความซับซ้อนของตัวแปรซึ่งอาจมีน้ำหนักต่อปัจจัยหลายปัจจัย

การวิเคราะห์ปัจจัย (Factor Analysis) (สุชาติ ประสิทธิ์รัฐสินธุ์ และ ลัดดาวัลย์ ยอดมณี.,2527)

การวิเคราะห์ปัจจัยเป็นเทคนิคการจัดกลุ่มของตัวแปร ซึ่งเกิดขึ้นจากความสัมพันธ์ระหว่างกันและกันของตัวแปรทำให้ทราบถึงโครงสร้างและแบบแผนของข้อมูล และหาปัจจัยร่วมของตัวแปรได้กล่าวคือเมื่อผู้วิจัยมีจำนวนตัวแปรมากๆ หลายตัว และมีความไม่สะดวกในการที่จะใช้ตัวแปรจำนวนมากดังกล่าวมาวิเคราะห์ เทคนิคการวิเคราะห์ปัจจัยจะลดจำนวนตัวแปรเหล่านั้นให้เหลือน้อยตัวโดยอาศัยโครงสร้างและแบบแผน (Structure and Pattern of Data) ของความสัมพันธ์ที่มีอยู่ในข้อมูลหรือระหว่างตัวแปร ตัวแปรที่นำมาวิเคราะห์เป็นตัวแปรเชิงปริมาณที่มีการแจกแจงแบบปกติ หรืออาจมีตัวแปรหุ่น (Dummy Variable) ได้บ้าง

การวิเคราะห์ปัจจัยยึดหลักที่ว่าการที่ตัวแปรหรือข้อมูลต่างๆ มีความสัมพันธ์กันก็เพราะตัวแปรต่างๆ เหล่านี้มีปัจจัยร่วมกัน (common factors) สังเกตได้จากการจับกลุ่มของตัวแปรหรือค่าสัมประสิทธิ์ความสัมพันธ์ระหว่างตัวแปร สมมุติว่ามีตัวแปร 20 ตัว และตัวแปรเหล่านี้มีความสัมพันธ์กันแบ่งออกได้เป็น 2 กลุ่มหรือ 2 ปัจจัย ในแต่ละกลุ่มตัวแปรจะมีความสัมพันธ์กันสูง การที่เป็นเช่นนี้ก็เพราะว่าตัวแปรเหล่านี้มีปัจจัยร่วมกัน ถ้าพบว่าปัจจัยร่วมและตัวแปรเหล่านี้มีความสัมพันธ์กันสูง แทนที่จะใช้ตัวแปรจำนวนมากๆ อาจใช้ปัจจัยร่วมแทนตัวแปรเหล่านี้ได้ เป็นการลดจำนวนตัวแปรให้น้อยลง

หลังจากที่หาปัจจัยร่วมของแต่ละกลุ่มได้แล้ว ยังสามารถที่จะหาคะแนนของแต่ละปัจจัยได้จากค่าของตัวแปรและอัตราของความสัมพันธ์ระหว่างตัวแปรกับปัจจัยร่วมแต่ละปัจจัย และสามารถนำคะแนนปัจจัยเหล่านี้ไปวิเคราะห์เพื่อศึกษาต่อเพิ่มเติมได้

การหมุนปัจจัย (วิยะดา ตันวัฒนากูล ,2542)

เป็นการแปลงเมทริกซ์เบื้องต้นให้เป็นเมทริกซ์ปัจจัย(Factor Transformation Matrix) ที่ง่ายต่อการตีความและการเข้าใจ การหมุนปัจจัยทำให้ตัวแปรบางตัวซึ่งแต่เดิมเป็นสมาชิกของหลายปัจจัยกลายเป็นสมาชิกของปัจจัยใดปัจจัยหนึ่งอย่างเด่นชัดขึ้นมากกว่าเดิม การที่ตัวแปรใดเป็นสมาชิกของปัจจัยใดจะดูจากน้ำหนักปัจจัยของตัวแปรนั้น ถ้าตัวแปรนั้นมีน้ำหนักปัจจัยบนหลายปัจจัยจะทำให้ยากต่อการตีความหรือระบุว่าตัวแปรนั้นเป็นสมาชิกของปัจจัยใด น้ำหนักปัจจัยของตัวแปรที่มีค่ามากที่สุดอยู่ในปัจจัยใด จะจัดว่าเป็นตัวแปรที่มีปัจจัยนั้นมาก

วิธีการหมุนปัจจัย

1.Varimax เป็นวิธีการหมุนปัจจัยแบบมุมฉาก (Orthogonal) โดยพยายามลดจำนวนตัวแปรที่มีน้ำหนักปัจจัยมากบนแต่ละปัจจัยให้เหลือน้อยที่สุด

2.Quartimax เป็นวิธีการหมุนปัจจัยแบบมุมฉาก (Orthogonal) ที่เน้นความง่ายในการตีความหมายของตัวแปร โดยพยายามหาตัวแปรให้น้อยที่สุดมาอธิบายตัวแปรแต่ละตัว

3.Equamax เป็นวิธีการหมุนปัจจัยแบบมุมฉาก (Orthogonal) ที่ผสมระหว่าง 2 วิธีข้างต้น

4.Oblimin เป็นวิธีการหมุนปัจจัยแบบมุมแหลม คือยอมให้ปัจจัยมีความสัมพันธ์กัน โดยวิธีนี้น้ำหนักปัจจัยและความสัมพันธ์ระหว่างปัจจัยและตัวแปรจะไม่เหมือนกัน

การวิเคราะห์องค์ประกอบหลัก (Galacitc Industries Corporation, 1985)

กระบวนการ PCA เป็นความเป็นไปได้ทางคณิตศาสตร์ที่จะลดชุดข้อมูลของการดูดกลืน Aij ของตัวอย่าง เมื่อข้อมูลถูกนำเข้ากระบวนการเต็มรูปแบบโดยใช้ PCA จะลดลงเหลือเมทริกซ์หลักเพียง 2 เมทริกซ์ คือ เมทริกซ์ของปัจจัยร่วม (common factor) และเมทริกซ์ของน้ำหนักปัจจัยหรือเมทริกซ์สหสัมพันธ์ระหว่างปัจจัยกับตัวแปร(factor loading) แสดงสมการโมเดลดังต่อไปนี้

$$\underset{(p\,x\,n)}{A} = \underset{(p\,x\,n)}{\mu} + \underset{(p\,x\,m)(m\,x\,n)}{LF} + \underset{(p\,x\,n)}{\varepsilon}$$

เมื่อ A คือ เมทริกซ์ของการดูดซับสเปกตรัมขนาด p x n

μ คือ ค่าเฉลี่ยของการดูดกลืนที่ความยาวคลื่นต่างๆ

L คือ เมทริกซ์ของน้ำหนักปัจจัยขนาด p x m

λ_{ij} =น้ำหนักปัจจัย หรือสหสัมพันธ์ระหว่างตัวแปรที่ i ปัจจัยร่วมที่ j

i = 1,2,...p และ j = 1,2,...m

F คือ เมทริกซ์ของปัจจัยร่วมขนาด m x n,

ε คือ เมทริกซ์ ความคลาดเคลื่อนของความสามารถการทำนาย calibration การดูดซับและมีมิติเหมือนกับ เมทริกซ์ A ในกรณีของการวิเคราะห์ eigenvectors มักจะเรียก เมทริกซ์ ε ว่าเป็น เมทริกซ์ ของสเปกตรัม residual eigenvector หาได้จาก eigenvalue

eigenvalue เป็นค่าการผันแปรรวมของตัวแปรทั้งหมดที่อธิบายได้โดยแกนแต่ละแกน

n คือ จำนวนของตัวอย่าง (สเปกตรัม)

p คือ จำนวนของชุดข้อมูล(ความยาวคลื่น)ที่ใช้ในการ calibration หรือจำนวนตัวแปร

m คือ จำนวนของปัจจัยร่วม

โดยมีข้อสมมติว่า(Johnson R. A. and Wichern D. W. ,1992)

F และ ε เป็นอิสระต่อกัน

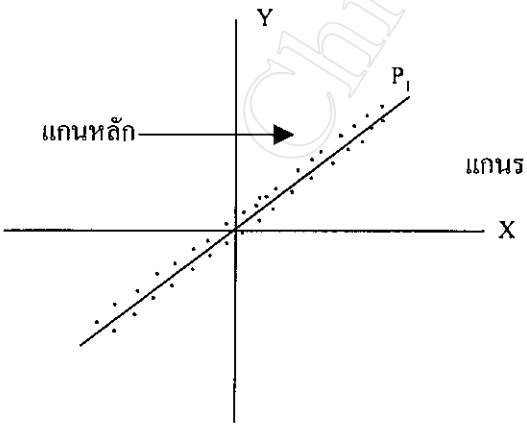
E(F) = 0 , Cov(F) = I (เมทริกซ์เอกลักษณ์)

$E(\varepsilon) = 0$,$Cov(\varepsilon) = \psi$ เมื่อ ψ เป็นเมทริกซ์ที่สมาชิกตัวอื่นๆ เป็น 0 ยกเว้นสมาชิกแนวทแยงมุม(diagonal matrix)

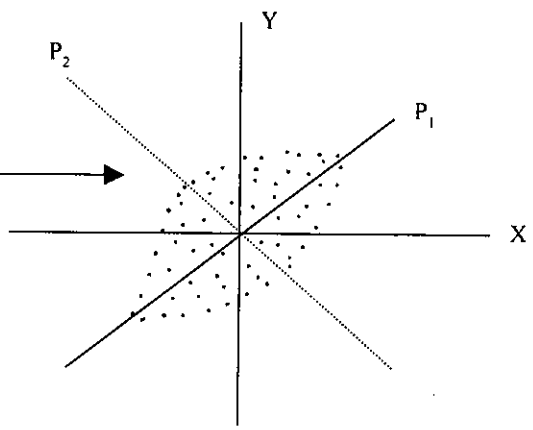
h_i^2 คือ ค่าความร่วมกันหรือ อัตราร่วมของปัจจัย (communality) เป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรหนึ่งกับตัวแปรอื่นๆ ทั้งหมด (R^2) ในปัจจัยนั้น ถ้าตัวแปรใดมีค่าความร่วมกันน้อยตัวแปรนั้นก็ควรถูกตัดออก (จากโปรแกรม SPSS ดูได้จาก Initial Statistics) สูตรการคำนวณคือ

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 \ldots + \ell_{im}^2$$

(สุชาติ ประสิทธิ์รัฐสินธุ์ และ ลัดดาวัลย์ ยอดมณี.,2527) การวิเคราะห์องค์ประกอบหลักเป็นวิธีการลดข้อมูล (ตัวแปร) ให้น้อยลงโดยอาศัยหลักความสัมพันธ์เชิงเส้นระหว่างตัวแปร (linear combination of the observed data) ที่ใช้เป็นข้อมูลแต่ไม่มีการสมมติเกี่ยวกับความสัมพันธ์เชิงสาเหตุและผลระหว่างปัจจัยและตัวแปร เช่น การวิเคราะห์องค์ประกอบหลักที่ใช้เมื่อมีตัวแปร 2 ตัว คือ x และ y ก่อนการวิเคราะห์องค์ประกอบหลักเริ่มจากการศึกษาความสัมพันธ์ระหว่างตัวแปร สมมุติว่า x และ y มีความสัมพันธ์กัน ซึ่งความสัมพันธ์กันนี้อาจดูได้จากการลงจุดบนกระดาษกราฟ ดังที่แสดงไว้ในภาพที่ 2 และ 3

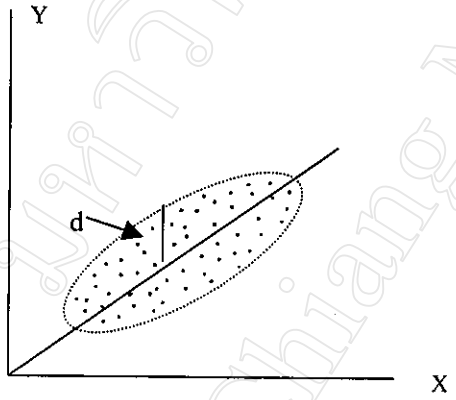


ภาพที่ 2 x และ y มีความสัมพันธ์กันสูง

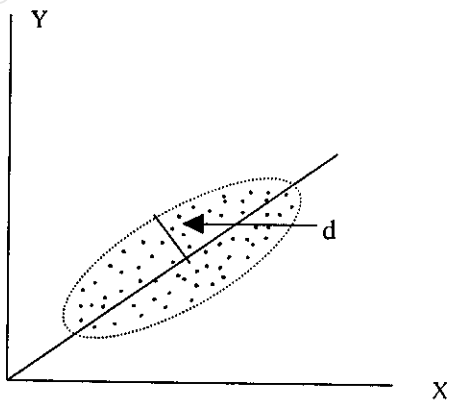


ภาพที่ 3 x และ y มีความสัมพันธ์กันต่ำ

ตามภาพที่ 2 x และ y มีความสัมพันธ์กันมาก และเป็นความสัมพันธ์เชิงเส้นตรงในทางบวก โดยที่ แกนหลัก (principal axis) คือ แกนที่สามารถกำหนดค่าของ x เมื่อทราบค่าของ y และกำหนดค่าของ y เมื่อทราบค่าของ x ถ้าสามารถกำหนดความลาดชันของเส้นตรงนั้นได้ เส้นนี้ก็คือ แกนหลัก ดังภาพที่ 2 ถ้าจุดต่างๆ อยู่บนเส้นแกนหลักทั้งหมด แกนหลักก็สามารถที่จะให้ข้อมูลเกี่ยวกับ x และ y ได้อย่างถูกต้องสมบูรณ์ทุกค่า แต่ถ้าจุดแสดงค่า x และ y กระจายออกไปมาก ต้องอาศัยแกนเพิ่มอีก 1 แกน ซึ่งแกนที่เพิ่มขึ้นนี้จะต้องมีจุดเริ่มต้นตั้งฉากกับแกนหลักดังภาพที่ 3 แกนหลักจะลากผ่านจุดต่างๆ ที่ทำให้ระยะระหว่างจุดกับแกนหลัก (โดยการลากเส้นจากจุดมาตั้งฉากกับแกนหลัก) สั้นที่สุด และทำให้ผลรวมของระยะทางยกกำลังสองมีค่าต่ำสุด การหาค่าต่ำสุดของแกนหลักนี้แตกต่างจากวิธีกำลังสองน้อยสุด (The least squares method) ซึ่งพยายามหาเส้นตรง Y=a+bx และให้ค่าที่ประมาณได้ Y' ต่างจาก Y เดิมน้อยที่สุด กล่าวคือ $(Y-Y')^2$ หรือ d_i^2 น้อยที่สุด การลากเส้นระยะทางตามแบบของวิธีการยกกำลังสองต่ำสุด เป็นการลากเส้นขนานกับแกน Y (แทนที่จะตั้งฉากกับแกนหลัก) ภาพที่ 4 และ 5 แสดงความแตกต่างระหว่างการหาค่าต่ำสุดของทั้ง 2 วิธี แม้ว่าทั้งสองวิธีจะพยายามทำให้ระยะทางต่ำสุดเช่นกัน (คือพยายามทำให้ $\sum d_i^2=0$)



ภาพที่ 4 เส้นระยะทางต่ำสุดของวิธีการยกค่ากำลังสองต่ำสุด



ภาพที่ 5 เส้นระยะทางต่ำสุดจากแกนหลักของวิธีการวิเคราะห์องค์ประกอบหลัก

ถ้ามีจำนวนตัวแปรเพิ่มขึ้นจำนวนมิติของแผนภาพจะเพิ่มขึ้น เช่นถ้ามี 3 ตัวแปร ต้องเพิ่มเส้นแสดงมิติอีก 1 เส้น และการลดจุดต้องคำนึงถึงค่าของตัวแปร 3 ตัว พร้อมๆ กัน และหาแกนหลักที่สามารถอธิบายการผันแปรของตัวแปรทั้ง 3 ตัวให้ได้มากที่สุด และแกนต่อๆ ไป เพื่ออธิบายการผันแปรที่เหลือให้ได้มากที่สุด

Johnson R. A. and Wichern D. W.(1992) ได้กล่าวว่า ในการวิเคราะห์ PCA โควาเรียนซ์เมทริกซ์ ของตัวอย่าง S จะจัดอยู่ในรูปคู่อันดับของ eigenvalue ($\hat{\lambda}_i$) และ eigenvector ($\hat{e}_i$) $(\hat{\lambda}_1,\hat{e}_1),(\hat{\lambda}_2,\hat{e}_2),...(\hat{\lambda}_p,\hat{e}_p)$ เมื่อ $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$ กำหนดให้ m < p จะสามารถประมาณค่าน้ำหนักของปัจจัย$\{\tilde{\lambda}_{ij}\}$ ได้จาก

$$\tilde{L} = [\sqrt{\hat{\lambda}_1}\hat{e}_1 \mid \sqrt{\hat{\lambda}_2}\hat{e}_2 \mid ... \mid \sqrt{\hat{\lambda}_m}\hat{e}_m]$$

เมื่อ p คือ จำนวนของชุดข้อมูล(ความยาวคลื่น)ที่ใช้ในการ calibration หรือจำนวนตัวแปร

m คือ จำนวนของปัจจัยร่วม

ประมาณค่าความร่วมกันของปัจจัยได้เท่ากับ $\tilde{h}_i^2 = \tilde{\lambda}_{i1}^2 + \tilde{\lambda}_{i2}^2 ... + \tilde{\lambda}_{im}^2$

สำหรับการวิเคราะห์ PCA โดยเมทริกซ์สัมประสิทธิ์สหสัมพันธ์สัมพันธ์ของตัวอย่าง (correlation matrix) จะใช้ เมทริกซ์สัมประสิทธิ์สหสัมพันธ์ แทน โควาเรียนซ์เมทริกซ์

ค่า eigenvalue ของปัจจัยตัวที่ j หาได้จาก

$$\hat{\lambda}_j = \lambda_{1j}^2 + \lambda_{2j}^2 + ... + \lambda_{pj}^2$$

สัดส่วนของการผันแปรรวมที่อธิบายได้โดยแกนแต่ละแกน (Proportion of total sample variance) $= \left\{\frac{\hat{\lambda}_j}{p}\right\}$

$$\% \text{ of variance} = \left\{\frac{\hat{\lambda}_j}{p}\right\} \times 100$$

ตัวอย่าง การวิเคราะห์องค์ประกอบหลักของ 5 ตัวแปรซึ่งมีเมทริกซ์สหสัมพันธ์ดังนี้

ตาราง1 เมทริกซ์สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร (correlation matrix)

ตัวแปร	V_1	V_2	V_3	V_4	V_5
V_1	1.00	0.02	0.96	0.42	0.01
V_2		1.00	0.13	0.71	0.85
V_3			1.00	0.50	0.11
V_4				1.00	0.79
V_5					1.00

หมายเหตุ ส่วนล่างของตารางเหมือนกับส่วนบน

จากตาราง 1 พบว่าตัวแปรที่มีความสัมพันธ์กันสูงสุดคือ V_1 กับ V_3 มีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.96 รองลงมาคือ V_2 กับ V_5 ค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.85 ทำการสกัดปัจจัยด้วยวิธี PCA ได้ค่าสถิติดังตาราง

ตาราง 2 ค่าสถิติจากการวิเคราะห์เมื่อสกัดปัจจัยได้ 2 ปัจจัย

ตัวแปร	น้ำหนักปัจจัย $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e}_i$		Communalities $\tilde{h}_i^2$
	F_1	F_2	
V_1	0.56	0.82	0.98
V_2	0.78	-0.53	0.88
V_3	0.65	0.75	0.98
V_4	0.94	-0.11	0.89
V_5	0.80	-0.54	0.93
Eigenvalue	2.85	1.81	
% of variance	57.00	36.20	
Cumulative %	57.00	93.20	

จากการสกัดปัจจัยด้วยวิธี PCA ได้ปัจจัย 2 ปัจจัย พิจารณาค่าน้ำหนักปัจจัยในแต่ละปัจจัย ตัวแปรใดมีค่าน้ำหนักปัจจัยในปัจจัยใดมากกว่า แสดงว่าตัวแปรนั้นควรอยู่ในปัจจัยนั้น และควรมีน้ำหนักปัจจัยในปัจจัยนั้นมากกว่า |0.3| จากตารางพบว่าปัจจัย F_1 มีค่าน้ำหนักปัจจัยของแต่ละตัวแปรมากกว่า 0.3 แสดงว่าปัจจัย F_1 ประกอบด้วยตัวแปร V_2,V_4และ V_5 ส่วนปัจจัย F_2 ประกอบด้วยตัวแปร V_1 และ V_3

ค่าความร่วมกันในตารางเป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรหนึ่งกับตัวแปรอื่นๆ ทั้งหมด ถ้ามีค่าต่ำตัวแปรนั้นก็ควรถูกตัดออก ซึ่งในตารางค่าความร่วมกันจะมีค่าสูง แสดงว่าไม่ควรตัดตัวแปรใดออกจากการวิเคราะห์

ค่า eigenvalue ของปัจจัย F_1 และ F_2 มีค่ามากกว่า 1 เป็นค่าการผันแปรรวมของตัวแปรทั้งหมดที่อธิบายได้โดยแต่ละปัจจัยแสดงว่าตัวแปรเหล่านี้แยกกลุ่มออกมาเป็น 2 ปัจจัยได้ ดี สัดส่วนของการผันแปรที่อธิบายได้ โดย F_1 เท่ากับ $\frac{2.85}{5} = 0.57$ หรือร้อยละ 57 ซึ่งแสดงว่าอัตราความสามารถของปัจจัย F_1สามารถอธิบายความผันแปรระหว่างตัวแปรทั้งหมดได้เท่ากับร้อยละ

57.00 ส่วนอัตราความสามารถของปัจจัย F_2 สามารถอธิบายความผันแปรระหว่างตัวแปรทั้งหมดได้ร้อยละ 36.20 และมีสัดส่วนความผันแปรสะสมของปัจจัย F_2 ได้ร้อยละ 93.20

การสร้างคะแนนปัจจัยหรือสเกลปัจจัย(สุชาติ ประสิทธิ์รัฐสินธุ์ และ ลัดดาวัลย์ ยอดมณี.,2527)

เนื่องจากคะแนนปัจจัยที่ได้จากการวิเคราะห์ปัจจัย เป็นคะแนนที่ได้จากค่าของตัวแปรต่างๆ หลายตัวที่รวมกลุ่มกันและมีปัจจัยร่วมกัน คะแนนของปัจจัยที่ได้จึงเปรียบเสมือนค่าของตัวแปรส่วนผสม (composite variable) โดยมีการให้น้ำหนักของตัวแปรแต่ละตัวตามน้ำหนักเชิงปัจจัยของตัวแปรนั้น

ประเภทของคะแนน

คะแนนที่สร้างได้จากการวิเคราะห์ปัจจัยสามารถจำแนกออกได้เป็น 2 ประเภท คือ

1.) คะแนนปัจจัย (Component scores) เป็นคะแนนที่สร้างจากน้ำหนักเชิงปัจจัยของตัวแปรทุกตัวที่มีปัจจัยร่วมกัน ที่ได้จากการวิเคราะห์ปัจจัย โดยอาศัยน้ำหนักของตัวแปรที่มีต่อแกนองค์ประกอบคูณกับค่าของตัวแปรในสมการดังตาราง 3

2.) สเกลที่มีฐานทางปัจจัย (Factor-based scale) เป็นคะแนนที่สร้างโดยอาศัยการวิเคราะห์ปัจจัยเป็นฐานแต่ไม่ได้เอาน้ำหนักเชิงปัจจัยของตัวแปรมาคิด เพียงแต่เอาค่าของตัวแปรที่มีน้ำหนักเชิงปัจจัยมากกว่า 0.3 มารวมกันเป็นคะแนน

โดยทั่วไปจะใช้คะแนนประเภทแรกเท่านั้น เช่น ในโปรแกรม SPSS for Windows เพราะมีการให้น้ำหนักต่างกันแก่ตัวแปรแต่ละตัว

วิธีการสร้างคะแนนปัจจัยด้วย การประมาณแบบถดถอย (regression estimates)

น้ำหนักเชิงปัจจัยที่จะใช้จากเมทริกซ์ค่าสัมประสิทธิ์คะแนนปัจจัย (Component score coefficient matrix) ในโปรแกรมสำเร็จรูป SPSS ดังตาราง 3 เป็นตารางตัวอย่างแสดงเมทริกซ์ค่าสัมประสิทธิ์คะแนนปัจจัยที่ใช้ในการหาคะแนนปัจจัย

ตาราง 3 เมทริกซ์ค่าสัมประสิทธิ์คะแนนปัจจัย (Standardized Component score coefficient matrix)

ตัวแปร	ปัจจัยที่ 1	ปัจจัยที่ 2
V_1	0.751	-0.322
V_2	0.323	0.002
V_3	0.212	0.637
V_4	0.301	0.581
V_5	0.004	0.782

คะแนนปัจจัยในแต่ละกรณี (case) ได้จากผลรวมของผลคูณระหว่างค่าสัมประสิทธิ์ของตัวแปรกับค่าของตัวแปร ซึ่งจะมี 2 ตัว เนื่องจากมี 2 ปัจจัย เมื่อ

$s_{1,i}$:คะแนนของปัจจัยที่ 1 ของกรณี ที่ i

$s_{2,i}$:คะแนนของปัจจัยที่ 2 ของกรณี ที่ i

Z_1 :คะแนนมาตรฐานของ V_1

Z_2 :คะแนนมาตรฐานของ V_2

Z_3 :คะแนนมาตรฐานของ V_3

Z_4 :คะแนนมาตรฐานของ V_4

Z_5 :คะแนนมาตรฐานของ V_5

$$s_{1,1} = 0.751Z_1 + 0.323Z_2 + 0.212Z_3 - 0.301z_4 + 0.00Z_5$$

$$s_{1,2} = -0.322Z_1 + 0.002Z_2 + 0.637Z_3 + 0.581Z_4 + 0.782Z_5$$

จะได้คะแนนปัจจัยในแต่ละกรณี เพื่อนำไปใช้ในการวิเคราะห์การถดถอยต่อไป

2.2 Inverse Least Square Regression

ผลการวิเคราะห์ PCA อย่างเดียวไม่สามารถนำมาใช้ทำนายความเข้มข้นของสารประกอบได้ จะต้องอาศัยเทคนิคการวิเคราะห์การถดถอยพหุคูณจาก ILS model ในการพยากรณ์หาค่าปริมาณน้ำตาล

2.2.1 Inverse Least Square (ILS) Model (บริษัทจาร์พา เท็คเซ็นเตอร์ จำกัด,2540)

เมื่อเลือกสารประกอบที่สนใจในตัวอย่างมา 1 ตัวอย่าง เช่น สารประกอบ C ภายในโปรตีน โปรตีนดูดกลืนแสงที่หลายๆความยาวคลื่นแต่ค่าคงที่ในการดูดกลืนแตกต่างกัน จากสมมติฐานที่ว่าค่าการดูดกลืนแสงที่แต่ละความยาวคลื่นมีความสัมพันธ์เชิงเส้นกับสารประกอบ C ภายในโปรตีนจะได้สมการ

$$P_1A_{11}+P_2A_{12}+P_3A_{13} = C$$

เมื่อ A_{ij} คือ การดูดกลืนของตัวอย่างที่ iที่ความยาวคลื่น j

P_j คือ ค่าสัมประสิทธิ์ของ A_{ij} สำหรับความยาวคลื่น j

สมมติให้ในที่นี้ใช้ความยาวคลื่น A_{ij} 3 ค่าที่ถูกเลือก คือ A_{11} , A_{12} และ A_{13}

ค่าคงที่ P_iมาจากกฎของ Beer-Lambert

$$A_i = \varepsilon_i l_i C_i$$

$$\frac{1}{\varepsilon_i l_i} A_i = C_i$$

ดังนั้น $$P_i = \frac{1}{\varepsilon_i l_i}$$

สำหรับตัวอย่างมาตรฐาน n ตัวอย่างจะมีกลุ่มสมการ

$$P_1A_{11}+P_2A_{12}+P_3A_{13} = C_1$$
$$P_1A_{21}+P_2A_{22}+P_3A_{23} = C_2$$
$$\vdots \qquad\qquad \vdots$$
$$P_1A_{n1}+P_2A_{n2}+P_3A_{n3} = C_n$$

ใช้ Multivariate Linear Least Square Regression เพื่อหาค่าที่เหมาะสมที่สุดของ Pi สำหรับตัวอย่างที่ไม่ทราบค่าที่มีค่าการดูดกลืน A_1*,A_2* และ A_3*ที่ความยาวคลื่น 3 ค่า สารประกอบ C *ภายในโปรตีนจะให้สมการดังนี้

$C^* = P_1A_1^* + P_2A_2^* + P_3A_3^*$

จะสามารถใช้วิธีการทำนองเดียวกันนี้ไปดัดแปลงตามแต่คุณสมบัติของตัวอย่าง เช่นไขมัน ความชื้น หรือ แป้ง ซึ่งข้อควรระวังคือการเลือกชุดความยาวคลื่น ที่เหมาะสมกับแต่ละคุณสมบัติของตัวอย่าง

สำหรับการวิจัยครั้งนี้ใช้คะแนนปัจจัย S จากการวิเคราะห์ PCA แทนที่ค่าการดูดกลืน A_{ij} ใน ILS โมเดล (Galacitc Industries Corporation,1985)

ดังนั้นสมการ โมเดลเป็นดังนี้:

$$\underset{(l \times n)}{C} = \underset{(l \times m)}{B}\underset{(m \times n)}{S} + \underset{(l \times n)}{E_c}$$

เมื่อ C คือ เมทริกซ์ ของความเข้มข้นของสารประกอบที่มีขนาด 1 x n

B คือ เมทริกซ์ 1 x m ของสัมประสิทธิ์การถดถอย

S คือ เมทริกซ์คะแนนปัจจัย จาก PCA model ขนาด m x n

E_c คือ เมทริกซ์ความคลาดเคลื่อนขนาด 1 x n

n คือ จำนวนตัวอย่าง (สเปกตรัม)

m คือ จำนวนปัจจัยร่วม

สัมประสิทธิ์ B สามารถหาได้โดยใช้วิธีวิเคราะห์การถดถอย

จะเห็นว่าการได้มาซึ่ง ILS โมเดลที่สมบูรณ์นั้นต้องใช้หลักการวิเคราะห์การถดถอยเข้ามาช่วย จึงขอเสนอหลักการและทฤษฎีในการวิเคราะห์การถดถอย ดังต่อไปนี้

2.2.2 การวิเคราะห์ถดถอยพหุคูณ

รูปแบบของสมการถดถอยเชิงซ้อน (กัลยา วานิชย์บัญชา, 2540)

กัลยา วานิชย์บัญชา. (2540) ถ้ามีตัวแปรอิสระ k ตัว $(X_1, X_2, ..., X_k)$ ที่มีความสัมพันธ์กับตัวแปรตาม Y โดยที่มีความสัมพันธ์อยู่ในรูปเชิงเส้น จะได้สมการถดถอยเชิงซ้อน ซึ่งแสดงความสัมพันธ์ระหว่าง Y และ $X_1, X_2, X_3, \ldots, X_k$ ดังนี้

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

โดยที่ β_0 = ส่วนตัดแกน Y เมื่อกำหนดให้ $X_1 = X_2 = \ldots = X_k = 0$

$\beta_1, \beta_2 x_1, \ldots, \beta_k$ เป็นสัมประสิทธิ์การถดถอย โดยที่ค่า β_i เป็นค่าที่แสดงถึงการเปลี่ยนแปลงของตัวแปรตาม Y เมื่อตัวแปรอิสระ X_i เปลี่ยนไป 1 หน่วย โดยที่ตัวแปรอิสระ X ตัวอื่นๆ มีค่าคงที่

ข้อกำหนดของการวิเคราะห์การถดถอย

1. ตัวแปรตามและตัวแปรอิสระมีการแจกแจงแบบปกติ
2. ตัวแปรอิสระเป็นตัวแปรหุ่น(Dummy Variable)ได้
3. ตัวแปรอิสระเป็นอิสระกัน

จากสมการถดถอย

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+e$$

การตรวจสอบเส้นสมการถดถอยที่เหมาะสม หลังจากได้เส้นแล้ว

1. ความคลาดเคลื่อน e เป็นตัวแปรที่มีการแจกแจงแบบปกติ
2. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์ นั่นคือ $E(e)=0$
3. ค่าแปรปรวนของความคลาดเคลื่อนเป็นค่าคงที่ $V(e)=V(y)$
4. e_i และ e_{i-1} เป็นอิสระต่อกัน

การประมาณค่าพารามิเตอร์ของสมการถดถอยเชิงซ้อน

จากสมการถดถอยเชิงซ้อน ซึ่งมีพารามิเตอร์ k+1 ตัว คือ $\beta_0,\beta_1,\beta_2,...,\beta_k$ การประมาณค่า $\beta_0,\beta_1,\beta_2,...,\beta_k$ จะต้องใช้ข้อมูลตัวอย่างของตัวแปร Y, X_1, X_2,...,X_k โดยใช้ตัวอย่างขนาด n จากสมการถดถอยเชิงซ้อน

จากกลุ่มประชากร

$$Y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+...+\beta_kx_{ki}+e_i \quad .....(2.1)$$

จะประมาณค่า Y_iของcase ที่ i หรือประมาณสมการที่ (2.1) ด้วย สมการที่ (2.2)

จากกลุ่มตัวอย่าง

$$\hat{Y}_i=b_0+b_1X_{1i}+b_2X_{2i}+...+bX_{ki} \quad .....(2.2)$$

ดังนั้นความคลาดเคลื่อนในการประมาณค่า Y_i ด้วย $\hat{Y}_i$ คือ $Y_i-\hat{Y}_i=e_i$ หรือเรียกว่า residual (สมการที่ (2.2) – (2.1))

การประมาณค่า $\beta_0,\beta_1,\beta_2,...,\beta_k$ ด้วยค่า $b_0,b_1,b_2,...,b_k$ ตามลำดับนั้นยังคงมีเป้าหมายเหมือนกับความถดถอยเชิงเส้นอย่างง่าย คือ เพื่อทำให้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองมีค่าน้อยที่สุด โดยใช้วิธีกำลังสองน้อยที่สุด นั่นคือหาค่า $b_0,b_1,b_2,...,b_k$ ที่ทำให้ $\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ มีค่าต่ำสุด

น้อยที่สุด โดยใช้วิธีกำลังสองน้อยที่สุด นั่นคือหาค่า $b_0,b_1,b_2,\ldots,b_k$ ที่ทำให้ $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ มีค่าต่ำสุด

ถ้า b_i เป็นลบแสดงว่า X_i กับ Y_i สัมพันธ์กันในทิศทางตรงกันข้าม ถ้า b_i เป็นบวกแสดงว่า X_i กับ Y_i สัมพันธ์กันในทิศทางเดียวกัน

จะเห็นได้ว่า X_i โดยวิธีการถดถอยอาจเทียบได้กับ A_{ij} ของวิธี Inverse Least Square นั่นเอง และ b_i ในวิธีการถดถอยอาจเทียบได้กับ P_i ของวิธี Inverse Least Square นั่นเอง

การทดสอบสมการความถดถอยเชิงซ้อนโดยใช้การวิเคราะห์ความแปรปรวน

จากสมการความถดถอยเชิงซ้อน

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

ค่าความแปรปรวนของ Y = ค่าแปรปรวนที่เกิดจากอิทธิพลของ X_1 , X_2 ,..., X_k + ค่าแปร ปรวนอย่างสุ่ม หรือ $SST = SSR + SSE$

โดยที่ SST (Sum Square of Total) คือค่าแปรปรวนทั้งหมดของ Y

หรือ $$SST = \sum_{i=1}^{n} \left(Y_i - \bar{y}\right)^2$$

SSR (Sum Square of Regression) คือ ค่าแปรปรวนของ Y เนื่องจากอิทธิพลของ $X_1,\ldots, X_k$

SSE (Sum Square of Error) หรือ ค่าแปรปรวนของ Y เนื่องจากอิทธิพลอื่นๆ หรือเรียกว่าค่าแปรปรวนอย่างสุ่ม

ตาราง 4 แสดงการวิเคราะห์ความแปรปรวนของการวิเคราะห์ความถดถอยเชิงซ้อน

แหล่งแปรปรวน (SV)	องศาอิสระ (DF)	ผลบวกกำลังสอง (SS)	ผลบวกกำลังสองเฉลี่ย(MS)	F
ความถดถอย (Regression)	k	SSR	MSR=SSR/k	$\frac{MSR}{MSE}$
ความคลาดเคลื่อน (Error)	n-k-1	SSE	MSE=SSE/(n-k-1)	
ผลรวม(Total)	n-1	SST		

โดยที่ $$SSR = \underline{b}'X'\underline{Y} - n\bar{y}^2$$

$$SST = \sum_{i=1}^{n}\left(Y_i - \bar{y}\right)^2 = \underline{Y}'\underline{Y} - n\bar{y}^2$$

$$SSE = \sum_{i=1}^{n}\left[Y_i - \left(a + b_1X_1 + b_2X_2 + ...,+b_kX_k\right)\right]^2$$

หรือ $SEE = SST - SSR = \underline{Y}'\underline{Y} - \underline{b}'X'\underline{Y}$

จากตารางการวิเคราะห์ความแปรปรวน จะใช้ในการทดสอบสมมติฐานเกี่ยวกับความสัมพันธ์ระหว่าง Y และ $X_1, X_2,..., X_k$ โดยตั้งสมมติฐานไว้ดังนี้

$$H_0 : \beta_1 = \beta_2 = ...\beta_k = 0$$

H_1 : มี β_i อย่างน้อย 1 ค่าที่ $\neq 0$; i = 1, 2, ...,k

สถิติทดสอบ $F = \dfrac{MSR}{MSE} = \dfrac{\left(\underline{b}'X'\underline{Y} - n\bar{y}^2\right)/k}{\left(\underline{Y}'\underline{Y} - \underline{b}'X'\underline{Y}'\right)/(n-k-1)}$

เขตปฏิเสธ จะปฏิเสธสมมติฐาน H_0 ถ้า $F > F_{k,n-k-1;1-\alpha}$ หรือ

ค่า sig F < α (เมื่อใช้โปรแกรม SPSS คำนวณ)

ผลของการทดสอบสมมติฐาน

ก. ถ้ายอมรับสมมติฐาน $H_0 : \beta_1 = \beta_2 = ...\beta_k = 0$ สรุปได้ว่า Y ไม่มีความสัมพันธ์กับ X ทั้ง k ตัว ($X_1, X_2,..., X_k$ ในรูปเชิงเส้น)

ข. ถ้าปฏิเสธสมมติฐาน H_0 หรือยอมรับสมมติฐาน H_1 อาจสรุปได้ว่ามี B_i อย่างน้อย 1 ตัวที่มีความสัมพันธ์กับ Y ในรูปเชิงเส้น จึงต้องทดสอบต่อไปว่า X_i ตัวใดที่มีความสัมพันธ์กับ Y โดยใช้ตัวสถิติทดสอบ t

โดยการทดสอบสมมติฐาน ดังต่อไปนี้

สมมติฐาน $H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$; i = 1, 2, ...,k

สถิติทดสอบ

$$t = \frac{b_i - \beta_i}{S_{b_i}} = \frac{b_i - 0}{S_{b_i}} = \frac{b_i}{S_{b_i}}$$

จะปฏิเสธสมมติฐาน H_0 เมื่อ $t > t_{1-\alpha/2\,;\,n-k-1}$ หรือ $t < -t_{1-\alpha/2\,;\,n-k-1}$ หรือกล่าวว่า

จะปฏิเสธ H_0 ถ้า $|t| > t_{1-\alpha/2\,;\,n-k-1}$ หรือปฏิเสธ H_0 เมื่อค่า sig t $\leq \alpha$ แสดงว่าตัวแปรนี้สมควรอยู่ในสมการ

ถ้า sig $> \alpha$ จะยอมรับ H_0 แสดงว่าตัวแปรนี้ไม่สมควรอยู่ในสมการ

การประมาณค่าคลาดเคลื่อนมาตรฐานของการถดถอย (Estimation of Standard Deviation of Regression)

การประมาณค่าคลาดเคลื่อนมาตรฐานของการถดถอย หรือเรียกกันว่าการประมาณค่าแปรปรวนของการประมาณความคลาดเคลื่อน ซึ่งเกิดจากการพยากรณ์หรือประมาณค่า Y ด้วย $\hat{Y}$ คือ e ในกรณีที่มีตัวแปรอิสระ k ตัว จะได้ค่าความแปรแปรวนของการประมาณคือ

ค่าคลาดเคลื่อนคือ Residual $= Y - \hat{Y}$

$$S_e^2 = S^2$$

โดยที่ $$S^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}$$

ดังนั้น ความคลาดเคลื่อนหรือค่าเบี่ยงเบนมาตรฐานของค่าประมาณ (Standard Error of Estimation : SE)

$$SE = \sqrt{S^2} = \sqrt{SSE/(n-k-1)} = \sqrt{MSE}$$

วิธีคัดเลือกตัวแปรเข้าสมการถดถอย (ต่าย เชี่ยงฉี่. ,2525)

มีหลายวิธีดังนี้

1. วิธี all possible regressions
2. วิธี backward elimination
3. วิธี forward selection
4. วิธี stepwise regression

1.วิธี all possible regressions

การพิจารณาหาสมการการถดถอย โดยวิธี all possible regressions คือให้ตัวแปรอิสระทุกตัวเข้าสมการ เพื่อพยากรณ์ตัวแปรตาม เช่น มีตัวแปรอิสระ x_i จำนวน k ตัว (i = 0, 1,..., k)

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0 ; i = 1, 2, \ldots, k$

วิธีนี้ให้พิจารณา ค่า sig t เพื่อปฏิเสธ หรือ ยอมรับ H_0 เอง

การพิจารณาหาสมการที่ดีที่สุดโดยวิธีนี้ จำเป็นต้องอาศัยประสบการณ์และความชำนาญของผู้ตัดสินใจอย่างมาก

2.วิธี backward elimination

หลักการของวิธี backward elimination ก็คือให้ตัวแปรอิสระทุกตัวเข้าสมการ และถูกตัดออกทีละตัวเมื่อตัวแปรนั้นไม่อยู่ในนัยสำคัญ

จากสมการถดถอยที่มีตัวแปรอิสระ x_i ทุกตัวในสมการและทดสอบ

$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0$ เมื่อผลการทดสอบได้ว่าปฏิเสธ H_0 จะพิจารณาตัดตัวแปร x_i ออกจากสมการ

3.วิธี forward selection

หลักการที่ใช้ในการเลือกสมการถดถอยโดยวิธี forward selection ก็คือพยายามเลือกตัวแปรอิสระ x_i ที่มีความสัมพันธ์กับตัวแปรตาม Y มากที่สุด เข้าไปในสมการก่อน มีขั้นตอนดังนี้

เลือกตัวแปร x_i ที่มีความสัมพันธ์กับตัวแปรตาม Y มากที่สุดเข้าไปในสมการถดถอยก่อน แล้วทดสอบ H_0: $\beta_i = 0$ ถ้าผลการทดสอบได้ว่าปฏิเสธ H_0 จะพิจารณาเพิ่มตัวแปร x_i เข้าไปในสมการอีก ถ้าผลการทดสอบได้ว่ายอมรับ H_0 ก็แสดงว่าตัวแปรขั้นตอนที่ 2 นี้จะไม่ถูกนำไปใส่ในสมการ ซึ่งก็จะได้สมการการถดถอยมีเฉพาะตัวแปร x_i ตามขั้นตอนที่ 1 อยู่ในสมการเท่านั้น แต่ถ้าผลการทดสอบได้ว่าปฏิเสธ H_0 ก็จะพิจารณานำตัวแปร x_i ตัวใหม่ใส่เข้าไปในสมการอีก โดยการทำซ้ำขั้นตอนที่ 2 ไปเรื่อยๆ จนได้สมการถดถอยที่ต้องการ

4.วิธี stepwise regression

การเลือกสมการถดถอย โดยวิธี วิธี stepwise regression จะเป็นการผสมผสานระหว่างวิธี backward elimination กับวิธี forward selection แล้วปรับปรุงให้ดีขึ้น มีขั้นตอนสรุปได้ดังนี้

1. นำตัวแปร x_i ที่มีความสัมพันธ์กับตัวแปรตาม Y มากที่สุดเข้าไปในสมการก่อน แล้วทดสอบ H_0: $\beta_i = 0$

2. เมื่อทำการทดสอบ H_0: $\beta_i = 0$ ในขั้นตอนที่ 1 พบว่าปฏิเสธ H0 ก็จะพิจารณาตัวแปร x_i ที่มีค่าสัมประสิทธิ์สหสัมพันธ์เชิงส่วนมีค่ามากที่สุดนำเข้าสมการแล้วทดสอบ H_0: all $\beta_i = 0$ และแต่

ละ H_0: β_i =0 ถ้าผลการทดสอบได้ว่ายอมรับ H_0 ก็จะนำตัวแปร x_i นั้นออกจากสมการ แต่ถ้าทดสอบได้ว่าปฏิเสธ H0 ก็จะพิจารณาเพิ่มตัวแปร x_i ที่เหลือเข้าไปใหม่

3. ดำเนินการตามขั้นตอนที่ 2 จนได้สมการถดถอยที่ต้องการ

สัมประสิทธิ์การตัดสินใจเชิงซ้อน (Multiple Coefficient of Determination : R^2)

กัลยา วานิชย์บัญชา. (2540) กล่าวว่าสัมประสิทธิ์การตัดสินใจเชิงซ้อนจะมีความหมายเหมือนกับความหมายของสัมประสิทธิ์การตัดสินใจ คือเป็นสัดส่วนหรือเปอร์เซ็นต์ที่ตัวแปรอิสระ (X_1 , X_2 , ..., X_k) สามารถอธิบายการเปลี่ยนแปลงของตัวแปรตาม Y ได้ หรืออาจกล่าวได้ว่าสัมประสิทธิ์การตัดสินใจเชิงซ้อนเป็นสัดส่วนหรือเปอร์เซ็นต์ของความแปรผันของตัวแปรตาม Y ที่มีสาเหตุเนื่องจากความผันแปรของ X_1 , X_2 , ..., และ X_k โดยที่สัมประสิทธิ์การตัดสินใจเชิงซ้อนจะให้สัญลักษณ์ $R^2_{Y.123...k}$ แต่โดยทั่วไปจะใช้ R^2

$$R^2 = \frac{\text{ความผันแปรของ Y เนื่องจากอิทธิพลของ } X_1, X_2, ..., X_k}{\text{ความแปรผันทั้งหมด}}$$

$$= SSR/SST$$

หรือ $$\boxed{R^2 = (SST - SSE)/SST = 1 - SSE/SST} \qquad \text{.....(2.3)}$$

โดยที่ $0 \le R^2 \le 1$

ถ้าค่า R^2 ที่ใกล้ 1 จะหมายถึง X_1 , X_2 , ..., X_k มีความสัมพันธ์หรือมีอิทธิพลต่อ Y มาก แต่ถ้าค่า R^2 เข้าใกล้ศูนย์ หมายถึงค่า X_1, X_2,..., X_k มีความสัมพันธ์กับ Y น้อย ในลักษณะพยากรณ์ที่สร้างขึ้น

เนื่องจาก SSR จะเพิ่มขึ้นถ้าเพิ่มตัวแปรอิสระ เช่น เดิมมี X_1 และ X_2 ที่มีความสัมพันธ์กับ Y แต่ถ้าเพิ่มตัวแปรอิสระ X_3 เข้าในสมการความถดถอย จะได้ว่า

$$SSR(X_2, X_2, X_3) > SSR(X_1, X_2)$$

โดยที่ $SSR(X_2, X_2, X_3)$ หมายถึง SSR ของสมการความถดถอยที่มีตัวแปรอิสระ X_1, X_2, และ X_3

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

และ $SSR(X_1, X_2)$ หมายถึง SSR ของสมการความถดถอยที่มีตัวแปรอิสระ X_1 และ X_2

ดังนั้นเมื่อเพิ่มตัวแปรอิสระเข้าสมการความถดถอยจะทำให้ค่า R^2 มากขึ้นทั้งที่ตัวแปรอิสระ X ที่เพิ่มขึ้นอาจจะไม่มีความสัมพันธ์กับ Y เลยก็ได้ จึงสามารถดูได้จากค่า R^2 change แล้วตรวจสอบ sig F change เพื่อดูว่าตัวแปรนี้สมควรเพิ่มเข้าไปในสมการหรือไม่

โดยมีการทดสอบสมมติฐาน ดังต่อไปนี้

สมมติฐาน $H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0 ; i = 1, 2, \ldots, k$

ถ้าปฏิเสธสมมติฐาน H_0 (sig Fchange $\leq \alpha$) แสดงว่าตัวแปรอิสระที่เพิ่มขึ้นนั้นสมควรเพิ่มในสมการ

นอกจากนี้ยังมีการปรับค่า R^2 ให้ถูกต้อง เรียกว่า Adjusted R^2 โดยที่

$$R_a^2 = \text{AdjustedR}^2$$

$$\boxed{R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4)$$

หรือ $$R_a^2 = 1 - \frac{(n-1)}{(n-k-1)}\left(R^2 - 1\right)$$

R^2_a ที่ได้สามารถใช้เปรียบเทียบสมการพยากรณ์ต่างๆ สมการพยากรณ์ใดมีค่า R^2_a มากกว่าแสดงว่า สมการนั้นเป็นสมการที่เหมาะสมกว่า

การพิจารณาเลือกสมการถดถอยที่ดีที่สุด นอกจากจะพิจารณาจากค่า R^2 และ ค่า R^2_a แล้วยังสามารถพิจารณาจากค่า PRESS (Prediction Residual Error Sum of Square) ซึ่งเป็นวิธีที่มีประสิทธิภาพมากที่สุดในกระบวนการวิเคราะห์ PCR อาจอธิบายได้ดังนี้

ทรงศิริ แต้สมบัติ. (2541).กล่าวว่า ค่า PRESS เป็นผลรวมกำลังสองของความคลาดเคลื่อนที่เกิดจากการประมาณค่าสังเกต Y_i ด้วยค่าประมาณ $\hat{Y}_{i(i)}$ ที่ได้จากสมการถดถอยจากข้อมูลตัวอย่างขนาด n-1 เมื่อไม่รวมค่าสังเกตที่ i

ค่าความคลาดเคลื่อน PRESS ที่ i = ค่าความคลาดเคลื่อนของ $\hat{Y}_{i(i)}$ จาก Y_i

$$= e_{i(i)}$$

$$= Y_i - \hat{Y}_{i(i)}$$

$$= \frac{Y_i - \hat{Y}_{i(i)}}{1 - \underline{X}_i (X'X)^{-1} \underline{X}'_i}$$

$$= \frac{e_i}{1 - h_{ii}}$$

ซึ่ง $\underline{X}_i$เป็นแถวนอนของเมทริกซ์ที่ I ของเมทริกซ์ X และ h_{ii}เป็นสมาชิกแนวเฉียงที่ i ของเมทริกซ์ $\underline{X}_i (X'X)^{-1} \underline{X}'_i$ จะมี

$$PRESS = \sum_{i=1}^{n} e_{i(i)}^2$$

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

รูปแบบการถดถอยที่เหมาะสมจะเป็นรูปแบบที่ให้ค่า PRESS น้อยที่สุด ซึ่งการคำนวณค่า PRESS ในการวิจัยครั้งนี้ได้มาจากการคำนวณผลรวมกำลังสองของค่าคลาดเคลื่อนที่ตัดออก $\sum d_i^2$

เมื่อค่าคลาดเคลื่อนของกรณีที่ i ถูกตัดออก เท่ากับ $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$

การตรวจสอบความเหมาะสมเส้นสมการถดถอยที่ได้อาจตรวจสอบได้ดังนี้

1. ค่าความคลาดเคลื่อนมีค่าเฉลี่ยเท่ากับ 0
2. ค่าคลาดเคลื่อนมีการแจกแจงแบบปกติ

โดยมี H_0: ค่าความคลาดเคลื่อนมีการแจกแจงแบปกติ

H_1:ค่าความคลาดเคลื่อนไม่มีการแจกแจงแบบปกติ

พิจารณาจากการทดสอบของ Kolmogorov และ Smirnov ถ้าค่า sig ของค่าสถิตินี้มีค่ามากกว่า α แสดงว่าค่าความคลาดเคลื่อนมีการแจกแจงแบบปกติ

3. ตรวจสอบความเป็นอิสระของค่าคลาดเคลื่อน

พิจารณาจากค่า Durbin-Watson ที่มีค่าใกล้ๆ 2 เป็นค่าที่แสดงว่าค่าความคลาดเคลื่อน (e_i กับ e_{i-1}) เป็นอิสระต่อกัน

การแก้ไขปัญหาค่าความคลาดเคลื่อนไม่เป็นอิสระต่อกันให้ปรับแก้ข้อมูลด้วยการหาผลต่าง (difference) ของตัวแปร โดยที่

$$A'_{ij} = A_{i,j} - A_{i-1,j}$$

เมื่อ A'_{ij} : ค่าการดูดกลืนที่ปรับแล้ว ของตัวอย่างที่ i ที่ความยาวคลื่น j

$A_{i-1,j}$: ค่าการดูดกลืนของตัวอย่างที่ i-1 ที่ความยาวคลื่น j

A_{ij} : ค่าการดูดกลืนของตัวอย่างที่ i ที่ความยาวคลื่น j

3.ตรวจสอบความคงที่ของค่าความแปรปรวนของค่าความคลาดเคลื่อน

โดยการสร้างแผนภาพการกระจายของค่าความคลาดเคลื่อน(e_i) กับค่าที่พยากรณ์ได้ ($\hat{Y}_i$) ถ้าหากแผนภาพที่ได้มีลักษณะกระจายไม่คงที่แสดงว่า ความแปรปรวนของค่าความคลาดเคลื่อนไม่คงที่ จะส่งผลให้การทดสอบสมมติฐานหรือการประมาณของช่วงพารามิเตอร์ในรูปแบบการถดถอยไม่ถูกต้อง การแก้ไขปัญหานี้อาจแก้โดยใช้วิธีกำลังสองน้อยที่สุดถ่วงน้ำหนัก (Weighted Least Square Method,WLS) ซึ่งในโปรแกรม SPSS คือ คำสั่ง Weight Estimation เพื่อหาตัวแปรที่จะใช้ถ่วงน้ำหนักตัวแปรตาม เพื่อนำไปถ่วงน้ำหนักในคำสั่ง Regression จะได้สมการถดถอยใหม่ที่ดีขึ้น

3.) สถิติวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างปริมาณน้ำตาลจากสมการcalibration กับปริมาณน้ำตาลที่วัดจากกรรมวิธีทางเคมี

ในการวิจัยครั้งนี้ใช้วิธีการทดสอบความสัมพันธ์ของค่าปริมาณน้ำตาลที่ได้จากการพยากรณ์กับค่าปริมาณน้ำตาลที่วัดได้จริงจากกรรมวิธีทางเคมี เพื่อดูว่าค่าปริมาณน้ำตาลที่พยากรณ์ได้มีค่าใกล้เคียงกับค่าปริมาณน้ำตาลที่วัดได้จริงมากน้อยเพียงใด โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ และหาค่าคลาดเคลื่อนมาตรฐานจากการพยากรณ์ (Standard Error of Prediction: SEP) ค่าความเอนเอียง (Bias) และค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variation : CV) ด้วย ดังต่อไปนี้

3.1 สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) (กัลยา วานิชย์บัญชา,2540)

เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ของตัวแปร 2 ตัว X และ Y ที่มีการแจกแจงปกติ และเป็นการทดสอบว่า X และ Y มีความสัมพันธ์กันหรือไม่

ให้ ρ เป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง 2 ตัวแปรของกลุ่มประชากร ในกรณีที่ค่าของ Y ขึ้นกับ X เพียงตัวเดียวจะเรียกว่า สัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation

Coefficient) โดยที่ ρ จะไม่มีหน่วย จึงสามารถใช้วัดความสัมพันธ์ระหว่าง X และ Y ได้ว่ามีความสัมพันธ์มากหรือน้อยเพียงใด เนื่องจากค่า ρ จะมีค่าสูงสุดเป็น 1 และต่ำสุดเป็น –1

ให้ r คือ สัมประสิทธิ์สหสัมพันธ์ของกลุ่มตัวอย่าง ซึ่งอาจให้ความหมายได้ดังนี้

1. ค่า r เป็นลบแสดงว่า X และ Y มีความสัมพันธ์ในทิศทางตรงข้าม กล่าวคือถ้า X เพิ่ม Y จะลด แต่ถ้า X ลด Y จะเพิ่ม

2. r เป็นบวกแสดงว่า X และ Yมีความสัมพันธ์ในทางเดียวกัน กล่าวคือถ้า X เพิ่ม Yจะเพิ่มด้วย แต่ถ้า X ลด Y จะลดด้วย

3. ถ้า r มีค่าเข้าใกล้ 1 หมายถึง X และ Y สัมพันธ์ในทิศทางเดียวกันและมีความสัมพันธ์กันอย่างสมบูรณ์

4. ถ้า r มีค่าเข้าใกล้ -1 หมายถึง X และ Y สัมพันธ์ในทิศทางตรงกันข้ามและมีความสัมพันธ์กันอย่างสมบูรณ์

5. ถ้า $|r|$ มีค่าน้อยกว่า 0.5 แสดงว่า X และ Yมีความสัมพันธ์กันน้อย

6. ถ้า r=0 แสดงว่า X และ Yไม่มีความสัมพันธ์กันในลักษณะสมการที่สร้างขึ้น

การทดสอบเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์

การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์ (ρ) เป็นการทดสอบว่า X และ Y มีความสัมพันธ์กันหรือไม่ โดยมีสมมติฐาน

$H_0: \rho = 0$ หรือ H_0: X และ Y ไม่มีความสัมพันธ์กันในรูปเส้นตรง

$H_1: \rho \neq 0$ H_0: X และ Y มีความสัมพันธ์กันในรูปเส้นตรง

สถิติทดสอบ

$$t = \frac{r - 0}{\sqrt{\hat{V}(r)}}$$

โดยที่ $\hat{V}(r) = \frac{1 - r^2}{n - 2}$

$\therefore$ ดังนั้น $$t = \frac{r}{\sqrt{\left(1 - r^2\right)/(n-2)}}$$

เขตปฏิเสธสมมติฐาน จะปฏิเสธ H_0 ถ้า $|t| > t_{1-\alpha/2;n-1}$

หรือ sig t< $\alpha/2$

3.2 ค่าความคลาดเคลื่อนมาตรฐานจากการพยากรณ์ (Standard Error of Prediction: SEP) (Irawan *et al* . ,1995).

เมื่อ y_i : ค่าปริมาณน้ำตาลที่ได้จากการวัดทางกรรมวิธีทางเคมี

$\bar{y}_i$: ค่าเฉลี่ยของปริมาณน้ำตาลที่ได้จากการวัดทางกรรมวิธีทางเคมี

$\hat{y}_i$: ค่าปริมาณน้ำตาลที่พยากรณ์ได้จากสมการ calibration

n : จำนวนตัวอย่าง

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} [(y_i - \hat{y}_i) - BIAS]^2}{n-1}}$$

3.3 ค่าความเอนเอียง (Bias)

$$BIAS = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n}$$

3.4 ค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variation : CV)

$$CV = \frac{SEP}{\bar{y}_i} \times 100\%$$

2. งานวิจัยที่เกี่ยวข้อง

Inazu (1986) สถานีทดลองการเกษตรฮอกไกโดแถลงว่าขณะนี้ได้เพาะข้าวพันธุ์ใหม่เพื่อปรับปรุงรสชาติของข้าวที่ปลูกในฮอกไกโด ในขั้นตอนการเลือกพันธุ์ข้าวนั้น ต้องทำการวิเคราะห์ตัวอย่างกว่า 10,000 ชุดทุกๆ ครึ่งปี โดยมีเน้นในเรื่องรสชาติของข้าว ซึ่งประกอบไปด้วย โปรตีน ไขมัน ความชื้น และแป้ง ด้วยวิธีทดสอบที่เคยทำมาซึ่งสามารถทดสอบตัวอย่างได้สูงสุด 2,000 ชุด; การจะทดสอบตัวอย่าง 10,000 ชุดจึงเป็นไปไม่ได้ในทางปฏิบัติ ดังนั้น NIR spectroscopy จึงเป็นความก้าวหน้าครั้งสำคัญที่ช่วยแก้ปัญหานี้ได้ สถานีทดลองการเกษตรฮอกไกโดประมาณความสามารถในการวิเคราะห์ของ NIR spectroscopy ว่าเร็วกว่าวิธีเดิมถึง 90 เท่า

Kawano *et al.* (1992).ได้ศึกษาถึงเทคโนโลยีการตรวจสอบน้ำตาลในผลลูกท้อ โดยวิธี NIR ได้สมการถดถอยเชิงเส้นพหุคูณในรูปแบบ $d^2\log(1/R)$ เมื่อ R คือค่าการสะท้อนแสงที่ความยาวคลื่น 906 878 870 และ 889 nm ได้ค่าคลาดเคลื่อนมาตรฐานของสมการ calibration (SEC) เท่ากับ 0.48 ^{o}Brix และค่าคลาดเคลื่อนมาตรฐานของการพยากรณ์ (SEP) เท่ากับ 0.50 ^{o}Brix พบว่าวิธีนี้สามารถตรวจสอบปริมาณน้ำตาลของลูกท้อได้

Irawan *et al.*(1995) .ศึกษาการตรวจสอบน้ำตาลและกรดในผลแอ็ปเปิ้ล ด้วยค่าการสะท้อนแสงจากวิธี NIR วิเคราะห์ด้วยวิธีถดถอยเชิงเส้นพหุคูณ โดยเลือกช่วงการสะท้อนแสงที่ความยาวคลื่น 1400 – 2000 nm ได้ความยาวคลื่น 5 ในสมการ calibration โดยได้ค่า r = 0.84 SEP=0.599 เมื่อพยากรณ์ ฟรุคโตส (n=20) ได้ค่า r=0.67 SEP =0.311 เมื่อพยากรณ์กลูโคส (n=20) r=0.60 SEP = 0.466เมื่อพยากรณ์ซูโครส (n=15) r=0.85 SEP =0.792 เมื่อพยากรณ์ฟรุคโตสรวมกลูโคส (n=20) r=0.74 SEP = 1.386 เมื่อพยากรณ์ฟรุคโตสรวมกลูโคสรวมกับซูโคส (n=20) ได้ค่า r=0.80 SEP =1.643 ^{o}Brix เมื่อพยากรณ์น้ำตาลรวม(n=30) และได้ค่า r=0.76 SEP=0.091 เมื่อพยากรณ์กรด(n=20)

Guthrie and Walsh (1997) ได้ศึกษา การประเมินคุณภาพของสับปะรดและมะม่วงโดยไม่บุบสลายโดยการใช้ Near Infrared Spectroscopy พบว่าการใช้ NIR ในการตรวจสอบคุณภาพของสับปะรดและมะม่วง โดยเลือกการสะท้อนกลับของคลื่นแม่เหล็กไฟฟ้าที่อยู่ในช่วง 760 - 2500 nm ค่าการดูดกลืนที่ความยาวคลื่นต่างๆ จะถูกสร้างความสัมพันธ์กับปริมาณความหวานน้ำสับปะรดและมะม่วงตากแห้ง(DM) เมื่อวิเคราะห์ถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression ,MLR) ของปริมาณความหวานของสับปะรด จะได้สมการมาตรฐานจากปริมาณสารที่

เป็นองค์ประกอบต่างๆความยาวคลื่น 866 ,760 ,1232 และ 832 nm ได้ค่า R^2 =0.75 และค่าความคลาดเคลื่อนมาตรฐานของการวัด (Standard Error of Calibration :SEC) = 1.21 ^{o}Brix (หน่วยความหวาน) เมื่อทำ Modified Partial Least Squares (MPLS) จะได้ค่า R^2 =0.91 ,SEC = 0.69 ^{o}Brix และ Standard Error of Cross Validation ,SECV = 1.09 ^{o}Brix ในมะม่วงจะทำ MLR โดยใช้คลื่นแม่เหล็กไฟฟ้าที่ 904,872,1660 และ 1516 nm ได้ค่า R^2 =0.90 และ SEC =0.85 % DM เมื่อทำ MPLS ได้ค่า R^2 =0.98 ,SEC = 0.54 Brix SECV = 1.19 พบว่าวิธีนี้สามารถตรวจสอบความหวานของสับปะรดและมะม่วงได้

Yantarasri *et al.* (1997) ได้ศึกษาการตรวจวัดน้ำตาลในผลมะม่วงโดยไม่ทำให้ตัวอย่างตรวจเสียหายพบว่าการตรวจวัดน้ำตาลในผลมะม่วงโดยไม่ทำให้ตัวอย่างตรวจเสียหายด้วยวิธี NIR โดยใช้คุณสมบัติของแสงเป็นตัวบอกปริมาณน้ำตาล สามารถพยากรณ์ความหวาน (Brix Value) ด้วยค่าสะท้อนแสงที่ความยาวคลื่นที่ 901 , 884 , 1060 และ 788 nm ได้สมการ

ปริมาณความหวาน (Brix value) = 20.17-927.60L(901)+679.81L(884)+386.24L(1060)

-251.73L(788)

L(λ)เป็นค่าฟังก์ชั่นในรูปของความสัมพันธ์ของปริมาณสารที่เป็นองค์ประกอบกับปริมาณความหวานตามความยาวคลื่นต่างๆ

λ= ความยาวคลื่น

โดยได้ค่า R^2 =0.95 และพบว่าสามารถตรวจสอบความหวานได้จริง