ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

ปรัชญาดุษฎีบัณฑิต (สถิติ)

ปริญญา

สถิติ | สถิติ

สาขา | **ภาควิชา**

เรื่อง การตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้องค์ประกอบรอง

Detection of Outliers for Multivariate Data Using the Minor Principal Components

นามผู้วิจัย นางรุ้งรวี อำนาจตระกูล

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(รองศาสตราจารย์อุษณีย์ ลีรวัฒน์, วท.ด.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(ผู้ช่วยศาสตราจารย์บุญอ้อม โฉมที, Ph.D.)

หัวหน้าภาควิชา .....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้องค์ประกอบรอง

Detection of Outliers for Multivariate Data Using the Minor Principal Components

โดย

นางรุ้งรวี อำนาจตระกูล

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาปรัชญาดุษฎีบัณฑิต (สถิติ)
พ.ศ. 2555

รุ้งรวี อำนาจตระกูล 2555: การตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้องค์ประกอบรอง ปริญญาปรัชญาดุษฎีบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: อาจารย์อำไพ ทองธีรภาพ, Ph.D. 87 หน้า

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้องค์ประกองรอง กรณีที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน โดยศึกษาหลักการ และทฤษฎีที่เกี่ยวข้องกับวิธีการวิเคราะห์องค์ประกอบหลัก รวมทั้งสมบัติการแจกแจงแบบปกติ ทำให้ได้ตัวสถิติ R_{2z}^2 และ R_{3z}^2 ซึ่งใช้เพียงองค์ประกอบรอง 2 และ 3 องค์ประกอบ ตามลำดับ พบว่า ตัวสถิติทั้งสองมีการแจกแจงแบบไคกำลังสอง ที่มีองศาอิสระเท่ากับ 1 เมื่อทำการเปรียบเทียบประสิทธิภาพของตัวสถิติ R_{2z}^2 และ R_{3z}^2 กับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanobis distance ในการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร จะทดลองกับข้อมูลจำลองภายใต้สถานการณ์ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร กำหนดจำนวนตัวแปร $p = 5, 10(2)20$ ขนาดข้อมูลตัวอย่าง $n = 30, 40, 50, 75, 100$ และร้อยละของค่าผิดปกติ 4 ระดับ ได้แก่ 10, 20, 30 และ 40 รวมทั้งหมด 140 สถานการณ์ ในแต่ละสถานการณ์ทำซ้ำจำนวน 1,000 ครั้ง

จากผลการศึกษา พบว่า ตัวสถิติ R_{2z}^2 จะเหมาะสำหรับตรวจสอบค่าผิดปกติเมื่อมีจำนวนตัวแปรตั้งแต่ 16 ตัวแปรขึ้นไป และขนาดข้อมูลตัวอย่างมากกว่า 50 ชุดข้อมูล แต่ตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าทุกตัวสถิติในหลายกรณี จึงยังไม่เหมาะที่จะนำไปใช้ในการตรวจสอบค่าผิดปกติ ส่วนตัวสถิติ $d_{2i,(2)}^2$ จะเหมาะสำหรับตรวจสอบค่าผิดปกติเมื่อมีจำนวนตัวแปรไม่เกิน 10 ตัวแปร แต่เมื่อมีจำนวนตัวแปรอยู่ระหว่าง 12 ถึง 16 ตัวแปร ตัวสถิติ $d_{2i,(3)}^2$ จะเหมาะสำหรับตรวจสอบค่าผิดปกติมากกว่า สำหรับตัวสถิติ d^2 จะเหมาะสำหรับตรวจสอบค่าผิดปกติเมื่อมีจำนวนตัวแปรตั้งแต่ 14 ตัวแปรขึ้นไป อย่างไรก็ตามทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน โดยตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีรูปแบบการคำนวณที่ไม่ยุ่งยาก สามารถคำนวณได้ง่าย และรวดเร็ว

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Rungrawee Amnarttrakul 2012: Detection of Outliers for Multivariate Data Using the Minor Principal Components. Doctor of Philosophy (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Mrs. Ampai Thongteeraparp, Ph.D. 87 pages.

The objective of this study is to propose test statistics to detect outliers for multivariate data using the minor principal components when the outliers do not conform with the correlation structure of the remainder of the data. After studying principals and theories on principal component analysis and properties of normal distribution, the test statistics R_{2z}^2 and R_{3z}^2 , which use only 2 and 3 minor principal components respectively, were found. The study shows that the test statistics are Chi-square distribution with degree of freedom equal to 1. The comparison of the efficiency of the test statistics for detection of outliers in multivariate data with $d_{2i,(2)}^2$ and $d_{2i,(3)}^2$, which are proposed by Hawkins, and d^2, which comes from Mahalanobis distance is tested by generating data through simulation technique using multivariate normal distribution under the condition of number of variables $p = 5$, 10(2)20 with a sample size of $n = 30$, 40, 50, 75, 100 and 4 levels of percentage of outliers 10%, 20%, 30% and 40%. There are 140 situations in total and each situation can be repeatedly simulated for 1,000 times.

The study shows that R_{2z}^2 is suitable for detection of outliers with $p \geq 16$ and $n \geq 50$. The efficient of R_{3z}^2 is lower than other test statistics in many situations, so it is not suitable to detect the outliers. The test statistic $d_{2i,(2)}^2$ is suitable for detection of the outliers with $p \leq 10$, whereas $d_{2i,(3)}^2$ is more suitable for detection of outliers with $12 \leq p \leq 16$. As for test statistic d^2 is suitable for detection of outliers with $p \geq 14$. However, it is discovered that all statistics are not much different, and it is found that R_{2z}^2 and R_{3z}^2 are not complicated and easily computed.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ดร.อำไพ ทองธีรภาพ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รศ.ดร. อุษณีย์ ลีรวัฒน์ และ ผศ.ดร. บุญอ้อม โฉมที อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ให้คำปรึกษาในการเรียน การค้นคว้าวิจัย ตลอดจนการตรวจแก้ไขวิทยานิพนธ์จนกระทั่งเสร็จสมบูรณ์ และกราบขอบพระคุณ รศ. ประสิทธิ์ พยัคฆพงษ์ ประธานการสอบ ผศ.ดร. กุศยา ปลั่งพงษ์พันธ์ ผู้ทรงคุณวุฒิภายนอก ที่ได้ให้ความกรุณาตรวจแก้ไขวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณอาจารย์ภาควิชาสถิติทุกท่าน ที่ได้อบรมสั่งสอนและมอบความรู้อันเป็นประโยชน์อย่างยิ่งในการนำไปใช้ประโยชน์ต่อไป

ด้วยความดีหรือประโยชน์อันใดเนื่องจากวิทยานิพนธ์เล่มนี้ ขอมอบแด่คุณพ่อ คุณแม่ ที่ได้อบรมและให้กำลังผู้วิจัยมาตลอดในทุกเรื่อง

รุ้งรวี อำนาจตระกูล

มกราคม 2555

สารบัญ

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

การตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้องค์ประกอบรอง

Detection of Outliers for Multivariate Data Using the Minor Principal Components

คำนำ

ในการวิเคราะห์ข้อมูลโดยใช้วิธีการทางสถิติ ขั้นตอนแรกคือ การเตรียมข้อมูลเพื่อนำเข้าสู่กระบวนการวิเคราะห์ สิ่งหนึ่งที่ผู้วิเคราะห์ควรคำนึงถึงคือ การตรวจสอบว่าข้อมูลมีค่าผิดปกติ (outlier) หรือไม่ ปัจจุบันมีผู้นิยามความหมายของค่าผิดปกติไว้มาก ความหมายของค่าผิดปกติที่ได้รับการอ้างถึงเป็นส่วนใหญ่ คือ Barnett and Lewis (1994) กล่าวว่าค่าผิดปกติหมายถึง ข้อมูลหรือกลุ่มของข้อมูลที่ไม่สอดคล้องกับกลุ่มของข้อมูลที่เหลืออยู่ และ Grubbs (1969) กล่าวว่าค่าผิดปกติเป็นค่าที่เบี่ยงเบนอย่างชัดเจนไปจากสมาชิกของกลุ่มตัวอย่าง โดยค่าผิดปกติที่ปะปนมากับข้อมูลนั้นสามารถตรวจสอบได้จากวิธีการที่แตกต่างออกไป สำหรับข้อมูลที่มีตัวแปรเดียว (univariate data) การตรวจสอบค่าผิดปกติสามารถทำได้ไม่ยาก วิธีการที่นิยมใช้กัน คือ วิธีทางกราฟ (graphical method) เป็นวิธีที่พิจารณาได้สะดวก รวดเร็ว และสามารถตรวจสอบค่าผิดปกติได้มากกว่า 1 ค่า แต่เป็นวิธีการพิจารณาที่ไม่เป็นทางการ (informal approach) ซึ่งการพิจารณาจะขึ้นอยู่กับแต่ละบุคคล และมีผู้เสนอให้นำวิธีทางกราฟไปปรับใช้กับการตรวจสอบค่าผิดปกติของข้อมูลหลายตัวแปร (Bacon-Shone and Fung, 1987)

ปัญหาในการตรวจสอบค่าผิดปกติของข้อมูลหลายตัวแปรทำได้ยาก กรณีข้อมูลมีจำนวนมาก และเมื่อจำนวนตัวแปรเพิ่มขึ้น วิธีการตรวจสอบจะยุ่งยากมากขึ้น (Rocke and Woodruff, 1996) ค่าผิดปกติของข้อมูลหลายตัวแปรจะไม่สามารถตรวจสอบได้จากแผนภาพแบบจุด (dot diagrams) ทีละตัวแปร หรือทีละคู่ วิธีที่นิยมใช้ในการตรวจสอบค่าผิดปกติของข้อมูลหลายตัวแปร คือเริ่มจากสร้างแผนภาพแบบจุดทีละตัวแปร จากนั้นสร้างแผนภาพการกระจาย (scatter plot) ของตัวแปรทีละคู่ และทำการคำนวณค่ามาตรฐาน (standardized values) ของข้อมูลทุกตัว โดยสังเกตข้อมูลที่มีค่ามาตรฐานที่ใหญ่หรือเล็ก จากนั้นคำนวณค่ากำลังสองของระยะห่างของข้อมูล (generalized squared distances) ข้อมูลที่ให้ค่ากำลังสองของระยะห่างของข้อมูลสูงมักจะเป็นค่าที่ผิดปกติ (Johnson and Wichern, 2007) แต่วิธีการเช่นนี้อาจเกิดข้อผิดพลาดได้ กล่าวคือ ผลการตรวจสอบพบว่าข้อมูลนั้นเป็นค่าผิดปกติ แต่ความจริงไม่ใช่ค่าผิดปกติ (swamping effect) หรือผลการตรวจสอบพบว่าข้อมูลนั้นไม่ผิดปกติ แต่ความจริงเป็นค่าผิดปกติ (masking effect) จึงมี

ผู้เสนอวิธีการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรโดยใช้หลักการต่างๆ เช่น Wilk (1963) เสนอวิธีการตรวจสอบค่าผิดปกติ จากข้อมูล k กลุ่มที่มีการแจกแจงแบบปกติ เมื่อไม่ทราบค่าพารามิเตอร์ โดยใช้ตัวสถิติ r_1 สำหรับการทดสอบค่าผิดปกติเพียง 1 ค่า และตัวสถิติ r_2 สำหรับการทดสอบค่าผิดปกติจำนวน 2 ค่า จากนั้นประยุกต์ตัวสถิติ r_1 และ r_2 เข้าด้วยกันนำไปสู่ตัวสถิติ r_t เพื่อทดสอบค่าผิดปกติจำนวน t ค่า Gnanadesikan and Kettenring (1972) เสนอแนวคิดและเทคนิคการตรวจสอบค่าผิดปกติ ด้วยวิธีการวิเคราะห์องค์ประกอบหลัก (principal component analysis) โดยพิจารณาจากตัวสถิติที่ใช้เฉพาะองค์ประกอบท้ายๆ หรือองค์ประกอบรอง (minor principal components) Rosner (1975) ได้เสนอวิธีการตรวจสอบค่าผิดปกติตั้งแต่ 1 ค่าขึ้นไป โดยใช้ตัวสถิติ R (RST) ซึ่งมีลักษณะเหมือนกับ Extreme Studentized Deviate (ESD) แต่คำนวณได้ง่ายกว่า จากการเปรียบเทียบอำนาจทดสอบพบว่า ตัวสถิติ R มีอำนาจการทดสอบดีกว่า Extreme Studentized Deviate (ESD) , Studentized Range (STR) และ Kurtosis (KUR) สำหรับการตรวจสอบค่าผิดปกติโดยใช้ Akaike's Information Criterion (AIC) ค่อนข้างแตกต่างจากวิธีที่กล่าวข้างต้น ถูกเสนอโดย Kitagawa (1979) AIC เป็นเกณฑ์ที่ใช้พิจารณาความเหมาะสมของตัวแบบ โดยจะสร้างตัวแบบเพื่อพยากรณ์ค่าผิดปกติ ซึ่งตัวแบบที่ดีจะต้องมีค่า AIC ต่ำสุด Schwager and Margolin (1982) ตรวจสอบค่าผิดปกติภายใต้แนวคิดที่ว่าตัวแบบของตัวอย่างสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร (multivariate normal distribution) จะแสดงว่าข้อมูลไม่ผิดปกติ แต่ถ้าข้อมูลผิดปกติจะไม่สอดคล้องกับตัวแบบ ในปี ค.ศ. 1984 Hawkins and Fatti ศึกษาการตรวจสอบหาค่าผิดปกติในข้อมูลหลายตัวแปร โดยใช้องค์ประกอบรองจากวิธีการวิเคราะห์องค์ประกอบหลัก และผลการตรวจสอบจะแม่นยำกว่าวิธีอื่น สำหรับ Formal Approach เป็นวิธีการตรวจสอบค่าผิดปกติที่ Munoz-Garcia *et. al.* (1990) ได้เสนอขึ้น ซึ่งใช้วิธีการวิเคราะห์ข้อมูลเชิงคุณภาพ โดยพิจารณาจากลักษณะของข้อมูล หรือความสัมพันธ์ระหว่างข้อมูล ข้อดีของวิธีการนี้คือตัวสถิติสามารถหาได้ไม่ยาก และไม่มีสมมติฐานเกี่ยวกับประชากร Hadi (1992) เสนอวิธีการหาค่าผิดปกติหลายค่า โดยพิจารณาจากระยะห่างของข้อมูลที่มีคุณสมบัติความแกร่ง (robustness) วิธีนี้มีประสิทธิภาพต่อปัญหา masking effect และ swamping effect ในปี ค.ศ. 1993 Davies and Gather เสนอวิธีการตรวจสอบค่าผิดปกติ โดยมีแนวคิดว่าค่าที่ผิดปกติจะต้องมีการแจกแจงที่ต่างไปจากข้อมูลตัวอื่นๆ และสร้างบริเวณของค่าผิดปกติ (outlier region) วิธีการนี้จะใช้ตัวสถิติทดสอบที่มีความแกร่ง และการทดสอบย้อนหลัง (outward testing) งานวิจัยของ Hadi (1994) ได้พัฒนาวิธีการหาค่าผิดปกติ และเปรียบเทียบอำนาจการทดสอบระหว่างวิธีการเดิมกับวิธีการใหม่ พบว่า วิธีการใหม่จะให้อำนาจการทดสอบที่ดีกว่า Rocke and Woodruff (1996) ได้เสนอวิธีการหาค่าผิดปกติภายใต้แนวคิดของ hybrid method ทำการพัฒนาขึ้นเป็น hybrid algorithm แบบใหม่ ซึ่งดีกว่าวิธีเดิม ใช้เวลาน้อย และสามารถตรวจสอบค่าผิดปกติได้จำนวนมาก

เมื่อข้อมูลมีจำนวนตัวแปรเพิ่มขึ้น ซึ่งการตรวจสอบค่าผิดปกติจะทำได้ยากขึ้น ในปี ค.ศ. 1999 Kosinski จึงสร้างวิธีการตรวจสอบค่าผิดปกติที่สามารถจัดการกับปัญหานี้ได้ดี โดยนำวิธี forward search มาประยุกต์ใช้ และศึกษาการเลือกระดับนัยสำคัญที่ทำให้พิจารณาได้ว่าข้อมูลใดจะเป็นค่าผิดปกติ ซึ่งเป็นวิธีที่ดีกว่าวิธีที่ Rocke and Woodruff (1996) ได้เสนอไว้ Caroni (2000) ได้ใช้วิธีการวิเคราะห์องค์ประกอบที่มีความแกร่ง (robust principal components analysis : RPCA) ทำการคำนวณค่าน้ำหนัก (weight) ของข้อมูลเพื่อนำไปประมาณค่าของแต่ละองค์ประกอบ และศึกษาการหาค่าวิกฤตสำหรับการทดสอบค่าผิดปกติจากค่าน้ำหนักของข้อมูลที่น้อยที่สุดใน RPCA ในปี ค.ศ. 2001 Pena and Prieto ได้เสนอวิธีการตรวจสอบค่าผิดปกติ ภายใต้แนวคิดของการวิเคราะห์ข้อมูลตัวแปรเดียวโดยพิจารณาจากทิศทาง (direction) ซึ่งกำหนดทิศทางโดยใช้ค่าสัมประสิทธิ์ของความโด่ง (kurtosis coefficient) เป็นวิธีที่ง่ายโดยเฉพาะเมื่อนำมาปรับใช้กับข้อมูลหลายตัวแปร Jackson and Chen (2004) นำวิธี RPCA มาใช้ตรวจสอบค่าผิดปกติกับข้อมูลทางนิเวศวิทยา โดยนำ minimum volume ellipsoid (MVE) มาปรับใช้กับวิธีการวิเคราะห์องค์ประกอบให้มีความแกร่งต่อค่าผิดปกติ ซึ่งให้ผลดีกว่าการตรวจสอบค่าผิดปกติจาก Mahalanobis distance กำลังสองที่คำนวณจากเมตริกซ์ความแปรปรวนร่วม Filzmoser *et. al.* (2008) เสนอวิธีตรวจสอบค่าผิดปกติที่สามารถทำได้รวดเร็ว และมีประสิทธิภาพกับข้อมูลที่มีตัวแปรจำนวนมาก ซึ่งจะใช้คุณสมบัติโดยทั่วไปของการวิเคราะห์องค์ประกอบหลัก และ robust kurtosis สามารถนำไปใช้กับข้อมูลจริงที่มีจำนวนตัวแปรมากกว่าจำนวนข้อมูลอย่างเช่นข้อมูลที่พบได้ในข้อมูลทางชีววิทยา

จากการศึกษาวิธีการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรดังกล่าวข้างต้น พบว่ามีแนวคิดแตกต่างกันออกไป สำหรับการวิเคราะห์องค์ประกอบหลักก็เป็นวิธีการหนึ่งที่มีผู้ทำการศึกษาแต่ยังไม่เป็นที่แพร่หลายมากนัก จุดประสงค์หลักที่แท้จริงของการวิเคราะห์องค์ประกอบหลัก คือ เพื่อลดจำนวนตัวแปรของข้อมูล เมื่อข้อมูลมีจำนวนตัวแปรมาก โดยให้ความสำคัญแก่องค์ประกอบแรกๆ หรือองค์ประกอบหลัก (major principal components) จำนวนน้อยที่สุดในการอธิบายรายละเอียดของข้อมูลเดิมให้ได้มากที่สุด โดยจะมีการตัดรายละเอียดบางส่วนของข้อมูลเดิมไป (Johnson and Wichern, 2007) การนำวิธีการวิเคราะห์องค์ประกอบหลักมาปรับใช้เพื่อตรวจสอบค่าผิดปกตินั้นสามารถทำได้ นอกจากการนำเฉพาะองค์ประกอบหลักมาใช้ในการตรวจสอบแล้วยังพบว่า องค์ประกอบรองนั้นยังสามารถนำมาใช้ในการตรวจสอบได้เช่นเดียวกัน แต่จะตรวจสอบความผิดปกติในลักษณะที่แตกต่างกัน กล่าวคือ องค์ประกอบหลักจะสามารถตรวจสอบค่าผิดปกติที่ปรากฏชัดเมื่อพิจารณาความผิดปกติของข้อมูลในแต่ละตัวแปร แต่สำหรับค่าที่ปกติในตัวแปรนั้นๆ เมื่อพิจารณารวมกับตัวแปรอื่น อาจจะพบว่าเป็นข้อมูลที่มี

ค่าผิดปกติก็ได้ นั่นคือตัวแปรไม่มีความสอดคล้องกัน เช่น เมื่อพิจารณาน้ำหนัก (กิโลกรัม) และส่วนสูง (เซนติเมตร) ของเด็กที่มีอายุในช่วง 5-15 ปี ถ้าเด็กคนหนึ่งมีน้ำหนัก 30 กิโลกรัม และส่วนสูง 175 เซนติเมตร จะพบว่าข้อมูลของเด็กคนนี้น่าจะผิดปกติ เนื่องจากตัวแปรทั้งสองไม่สอดคล้องกัน แต่ถ้าพิจารณาในแต่ละตัวแปร สามารถเป็นไปได้ที่เด็กในช่วงอายุดังกล่าวจะมีน้ำหนัก 30 กิโลกรัม และเมื่อพิจารณาส่วนสูงของเด็กในช่วงอายุนี้มีความเป็นไปได้ที่จะมีส่วนสูงถึง 175 เซนติเมตร แต่เป็นไปไม่ได้ที่เด็กจะมีน้ำหนักเพียง 30 กิโลกรัม และสูงถึง 175 เซนติเมตร ซึ่งความผิดปกติในลักษณะนี้จะไม่สามารถตรวจสอบได้จากองค์ประกอบหลัก แต่สามารถตรวจสอบได้จากองค์ประกอบรองเท่านั้น ดังที่ Jolliffe (2002) ได้กล่าวไว้ว่า ในการตรวจสอบค่าผิดปกติในลักษณะดังกล่าว อาจจะสามารถตรวจสอบโดยใช้ “the last few principal components” หรือองค์ประกอบ 2-3 องค์ประกอบสุดท้าย หรือองค์ประกอบรองนั่นเอง ซึ่งเป็นองค์ประกอบที่ผู้วิเคราะห์มักจะไม่ให้ความสำคัญ แต่อย่างไรก็ตามผู้ที่ให้ความสนใจและทำการศึกษาการตรวจสอบค่าผิดปกติโดยใช้องค์ประกอบรองรวมทั้งรูปแบบความผิดปกติที่เกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกันยังมีไม่มากนัก

ปัจจุบันวิธีการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรมีหลายวิธี วิธีที่นิยมใช้กันโดยทั่วไป ได้แก่ วิธี Mahalanobis distance ซึ่งเป็นวิธีที่พิจารณาจากความแตกต่างของข้อมูลกับค่ากลางของข้อมูลในรูปของค่ามาตรฐาน โดยข้อมูลที่มีความแตกต่างกับค่ากลางมากแสดงว่าอาจจะเป็นค่าที่ผิดปกติ ส่วนวิธีการตรวจสอบค่าผิดปกติที่ใช้แนวคิดของวิธีการวิเคราะห์องค์ประกอบหลักนั้นมีไม่มากนัก ซึ่งวิธีที่นำเสนอโดย Hawkins (1974) เป็นวิธีหนึ่งที่ให้ความสำคัญกับองค์ประกอบรองมากกว่าองค์ประกอบหลัก สามารถตรวจสอบค่าผิดปกติในลักษณะที่ไม่ปรากฏชัดในแต่ละตัวแปร หรือความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน ซึ่งเป็นความผิดปกติในลักษณะที่ตรวจสอบได้ยาก การนำองค์ประกอบรองจากวิธีการวิเคราะห์องค์ประกอบหลักมาใช้ในการตรวจสอบค่าผิดปกติจึงเป็นแนวคิดที่น่าสนใจศึกษา

การศึกษาวิจัยครั้งนี้จะทำการสร้างตัวสถิติขึ้นใหม่เพื่อใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรด้วยวิธีการวิเคราะห์องค์ประกอบหลักโดยใช้องค์ประกอบรอง ในลักษณะที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน และทำการเปรียบเทียบประสิทธิภาพวิธีการที่นำเสนอกับวิธีที่นำเสนอโดย Hawkins และวิธี Mahalanobis distance รวมทั้งทำการทดลองนำวิธีดังกล่าวไปปรับใช้กับข้อมูลจริง

วัตถุประสงค์

1. เพื่อสร้างตัวสถิติ R_{2z}^2 และ R_{3z}^2 ที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยใช้องค์ประกอบรอง กรณีที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรของตัวสถิติ R_{2z}^2 และ R_{3z}^2 จากวิธีที่นำเสนอ กับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จากวิธีที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanobis distance

การตรวจเอกสาร

การตรวจเอกสารแบ่งออกเป็น 2 ส่วน คือ งานวิจัยที่เกี่ยวข้อง และความรู้พื้นฐานที่เกี่ยวข้อง รายละเอียดมีดังนี้

งานวิจัยที่เกี่ยวข้อง

Rao (1964) ศึกษาเทคนิคของการวิเคราะห์ข้อมูลหลายตัวแปร โดยเน้นที่วิธีการวิเคราะห์องค์ประกอบหลักเพื่อนำผลการวิเคราะห์ที่ได้ไปปรับใช้กับงานวิจัยเชิงประยุกต์ ได้กล่าวถึงบทบาทที่สำคัญของวิธีการวิเคราะห์องค์ประกอบหลักในด้านต่างๆ และเป็นผู้ที่ได้เสนอตัวสถิติ d_{1i}^2 ซึ่งเป็นตัวสถิติทดสอบที่คำนวณได้จากผลรวมขององค์ประกอบรองยกกำลังสอง เพื่อใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร แต่ยังไม่มีการศึกษาว่าควรจะใช้องค์ประกอบรองจำนวนเท่าไรจึงจะเหมาะสม

Gnanadesikan and Kettenring (1972) ศึกษาแนวคิดและเทคนิคของวิธีการประมาณค่าวัดตำแหน่งและค่าวัดการกระจายที่มีความแกร่งสำหรับข้อมูลหลายตัวแปร การวิเคราะห์ ค่าเศษตกค้าง (residual) จากวิธีการวิเคราะห์องค์ประกอบที่สอดคล้องกับวิธีการวิเคราะห์กำลังสองน้อยที่สุด และการหาค่าผิดปกติสำหรับข้อมูลหลายตัวแปร ซึ่งการวิเคราะห์เหล่านี้กับข้อมูล หลายตัวแปรทำได้ยากแต่เป็นสิ่งที่จำเป็น โดยงานวิจัยนี้ทำการศึกษาตัวสถิติ d_{1i}^2 ของ Rao (1964) เพิ่มเติม ซึ่งได้ข้อสรุปว่าถ้าข้อมูลมีการแจกแจงแบบแกมมา แสดงว่าข้อมูลนั้นไม่มีค่าที่ผิดปกติ นอกจากนี้ยังได้เสนอตัวสถิติ d_{3i}^2 ที่สามารถใช้ในการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรที่ปรากฏชัดเจนในแต่ละตัวแปร ซึ่งจะให้ความสำคัญกับองค์ประกอบหลัก

Hawkins (1974) ทำการศึกษาตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูล หลายตัวแปร ในกรณีที่ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร ภายใต้หลักการของวิธีการวิเคราะห์องค์ประกอบหลัก โดยให้ความสำคัญกับองค์ประกอบที่มีค่าเฉพาะ (eigenvalues) ที่มี ค่าน้อยๆ หรือองค์ประกอบรอง และจากการศึกษาตัวสถิติ d_{1i}^2 ของ Rao (1964) พบว่า ในการคำนวณตัวสถิตินี้ควรจะคำนึงถึงน้ำหนักขององค์ประกอบรองด้วย เนื่องจากความแปรปรวนของแต่ละองค์ประกอบจะลดลงในองค์ประกอบท้ายๆ จึงได้เสนอให้มีการถ่วงน้ำหนัก และพัฒนาตัวสถิติขึ้นใหม่ คือ d_{2i}^2 โดยนำค่าเฉพาะมาคำนวณร่วมกับค่าองค์ประกอบรอง นอกจากนี้ยังได้เสนอตัวสถิติ d_{4i}^2 ซึ่งเป็น ตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติเช่นเดียวกัน แต่รูปแบบ

การคำนวณของค่าองค์ประกอบรองจะอยู่ในรูปของค่าสัมบูรณ์ และยังได้เสนอเกณฑ์ในการเลือกจำนวนองค์ประกอบรองที่ควรจะนำมาใช้กับตัวสถิติเหล่านี้ ซึ่งเป็นปัญหาที่ยังไม่มีข้อสรุปที่ชัดเจน

Campbell (1980) เสนอวิธีวิเคราะห์องค์ประกอบหลักที่มีความแกร่ง เพื่อตรวจสอบ ค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยพิจารณาด้วย probability plot ของ Mahalanobis distance ที่ใช้ M-estimator เป็นตัวประมาณค่าเฉลี่ยและความแปรปรวนร่วมแทนการใช้ ตัวประมาณภาวะความน่าจะเป็นสูงสุด (maximum likelihood estimator) และมีการคำนวณ ค่าน้ำหนักให้กับตัวประมาณที่มีความแกร่งนี้ เพื่อแสดงว่าข้อมูลตัวใดเป็นค่าผิดปกติ และพบว่าถ้าข้อมูลไม่มีค่าผิดปกติตัวประมาณที่แกร่ง (robust estimator) นี้จะมีลักษณะเช่นเดียวกับตัวประมาณภาวะความน่าจะเป็นสูงสุด

Hawkins and Fatti (1984) ทำการศึกษาวิธีการวิเคราะห์องค์ประกอบหลัก โดยให้ความสำคัญกับองค์ประกอบรอง โดยเสนอว่าองค์ประกอบรองสามารถถูกนำมาใช้ในการเลือกตัวแปรอิสระในการวิเคราะห์การถดถอยเชิงพหุ (multiple regression analysis) การพิจารณา ตัวแปรซ้ำซ้อน (redundant variable) และประการสุดท้ายคือ สามารถนำไปใช้ในการตรวจสอบ ค่าผิดปกติสำหรับข้อมูลหลายตัวแปร ด้วยการสร้างเมตริกซ์ D ซึ่งเป็นค่ามาตรฐานขององค์ประกอบมาใช้ในการพิจารณา

Marden (1999) กล่าวว่าข้อมูลหลายตัวแปรที่มีค่าผิดปกติจะส่งผลกระทบต่อเมตริกซ์ความแปรปรวนร่วม รวมถึงองค์ประกอบที่ได้จากวิธีการวิเคราะห์องค์ประกอบหลักด้วย จึงเสนอว่าการใช้เมตริกซ์ความแปรปรวนร่วมที่มีความแกร่งในวิธีการวิเคราะห์องค์ประกอบหลักจะสามารถแก้ปัญหานี้ได้ และได้เสนอหลักการ 2 แนวทาง คือ sign procedure และ rank procedure สำหรับข้อมูลหลายตัวแปร โดยใช้การแปลงข้อมูลให้อยู่ในเทอมของ signs และ ranks จากนั้นทำการหาเวกเตอร์เฉพาะ (eigenvectors) ของเมตริกซ์ความแปรปรวนร่วมที่ได้ทำการแปลงข้อมูลเหล่านี้ตามขั้นตอนเดิม ซึ่งงานวิจัยนี้ได้ทำการศึกษากรณี 2 ตัวแปร และกรณีที่ข้อมูลมีจำนวนตัวแปรมากขึ้นก็ยังคงใช้ได้ดี

Caroni (2000) เสนอวิธีการวิเคราะห์องค์ประกอบหลักที่มีความแกร่งที่พัฒนามาจากการศึกษาของ Campbell (1980) เพื่อตรวจสอบค่าผิดปกติ โดยจะทำให้ค่าประมาณของแต่ละองค์ประกอบให้มีความแกร่ง จากนั้นทำการคำนวณค่าน้ำหนัก ซึ่งค่าน้ำหนักนี้จะเป็นค่าที่ใช้ในการพิจารณาถึงค่าผิดปกติ นั่นคือ ค่าน้ำหนักน้อยจะแสดงว่าเป็นค่าที่ผิดปกติ และยังมีการคำนวณ

ค่าวิกฤต (critical value) สำหรับค่าน้ำหนักที่ใช้ในการตรวจสอบด้วยการจำลองสถานการณ์ (simulation)

Croux and Haesbroeck (2000) เสนอวิธีการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์ความแปรปรวนร่วม หรือเมตริกซ์สหสัมพันธ์ (correlation matrix) ที่มีความแกร่ง ในทางสถิติเมตริกซ์ทั้งสองชนิดนี้มีบทบาทสำคัญต่อเทคนิคของการวิเคราะห์ข้อมูลหลายตัวแปร เนื่องจากข้อมูลที่มีค่าผิดปกติจะส่งผลกระทบต่อการวิเคราะห์องค์ประกอบหลักแบบดั้งเดิม ซึ่งจะทำให้นำไปสู่การวิเคราะห์ที่ผิดพลาด จึงทำการเปรียบเทียบตัวประมาณที่มีความแกร่ง 3 ตัว ได้แก่ M-estimator , S-estimator และ one-step reweighted minimum covariance determinant estimator พบว่า ในทางปฏิบัติควรใช้ S-estimator แต่ถ้าไม่เน้นความสำคัญในเรื่องประสิทธิภาพและการอนุมานควรใช้ one-step reweighted minimum covariance determinant estimator

Shyu *et.al.*(2003) นำเสนอวิธีการตรวจสอบความผิดปกติของปัญหาการบุกรุก (instrusion detection) หรือข้อมูลที่มีค่าผิดปกติบนเครือข่ายคอมพิวเตอร์ โดยใช้วิธีการวิเคราะห์องค์ประกอบหลัก เริ่มจากการตัดข้อมูล (trimming) เพื่อหาตัวประมาณเมตริกซ์สหสัมพันธ์ที่มีความแกร่ง จากนั้นจึงทำการวิเคราะห์องค์ประกอบหลัก ซึ่งนำองค์ประกอบหลักที่สามารถอธิบายความผันแปรรวมกันได้ถึงร้อยละ 50 และองค์ประกอบรองที่มีค่าเฉพาะน้อยกว่า 0.20 มาพิจารณาร่วมกันเพื่อตรวจสอบหาค่าผิดปกติ จากการเปรียบเทียบประสิทธิภาพของวิธีนี้กับ วิธี density-based local outliers (LOF) วิธีพิจารณาจากระยะห่างโดยใช้ Canberra metric วิธีพิจารณาจากระยะห่างโดยใช้ Euclidean distance และวิธี k-nearest neighbor เมื่อ k=5 พบว่า วิธีใหม่นี้ให้ประสิทธิภาพดีที่สุด

Shan (2007) นำเสนอวิธีการตรวจสอบค่าผิดปกติด้วยวิธีวิเคราะห์องค์ประกอบหลัก และได้นำวิธีการปรับข้อมูลให้เรียบมาวิเคราะห์ร่วมด้วย ซึ่งปรับข้อมูลให้เรียบโดย Kernel function เพื่อเป็นการแปลงข้อมูลจากปริภูมิเดิม (original space) ไปสู่ปริภูมิใหม่ (feature space) จากนั้นนำข้อมูลดังกล่าวไปวิเคราะห์ด้วยวิธีวิเคราะห์องค์ประกอบหลัก โดยนำองค์ประกอบรองที่เล็กที่สุดมาใช้ในการพิจารณาค่าผิดปกติจากวิธีทางกราฟ ซึ่งเป็นวิธีที่ง่ายที่สุด

ความรู้พื้นฐานที่เกี่ยวข้อง

1. การแจกแจงแบบปกติหลายตัวแปร

การแจกแจงแบบปกติหลายตัวแปรคือ รูปทั่วไปของการแจกแจงแบบปกติของตัวแปรเดียวที่มีหลายมิติ (Evans *et. al.*, 2000) สำหรับการแจกแจงแบบปกติของตัวแปรเดียว ด้วยค่าเฉลี่ย μ และความแปรปรวน σ^2 มีฟังก์ชันความน่าจะเป็น (probability density function) คือ

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty<x<\infty \tag{1}$$

สามารถเขียนแทนด้วยสัญลักษณ์ $N(\mu,\sigma^2)$ ซึ่งจะขยายไปสู่ฟังก์ชันความน่าจะเป็นของการแจกแจงแบบปกติหลายตัวแปร

เมื่อพิจารณาเทอมที่เป็นส่วนยกกำลังของเลขชี้กำลัง e คือ

$$\left(\frac{x-\mu}{\sigma}\right)^2=(x-\mu)'(\sigma^2)^{-1}(x-\mu) \tag{2}$$

จากสมการ (2) จะแสดงถึงระยะห่างหรือความแตกต่างของ x และ μ ที่อยู่ในรูปมาตรฐานยกกำลังสอง ในทำนองเดียวกันสามารถขยายให้อยู่ในรูปทั่วไปของเวกเตอร์ $\mathbf{x}$ ที่มีขนาด $p\times 1$ ของข้อมูลที่มีหลายตัวแปร คือ

$$(\mathbf{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{3}$$

เช่นเดียวกับสมการ (3) แสดงถึงระยะห่างหรือความแตกต่างของ $\mathbf{x}$ และ $\boldsymbol{\mu}$ ที่อยู่ในรูปมาตรฐานยกกำลังสอง

โดยที่ $\boldsymbol{\mu}$ แทนค่าเฉลี่ยของเวกเตอร์ $\mathbf{x}$ ที่มีขนาด $p\times 1$

$\boldsymbol{\Sigma}$ แทนเมตริกซ์ความแปรปรวนร่วมของ $\mathbf{x}$ ที่มีขนาด $p\times p$

ฟังก์ชันความน่าจะเป็นของการแจกแจงแบบปกติหลายตัวแปรจะทำได้จากการเปลี่ยนสมการ (2) เป็นสมการ (3) ลงในสมการ (1) และเทอมค่าคงที่สำหรับตัวแปรเดียว (univariate normalizing constant) จะปรับเป็นค่าคงที่ที่ใช้สำหรับหลายตัวแปร ซึ่งอยู่ในรูป $1/(\sqrt{2\pi})^p \mid \mathbf{\Sigma} \mid^{1/2}$

ดังนั้นฟังก์ชันความน่าจะเป็นของการแจกแจงแบบปกติหลายตัวแปร $\mathbf{x} = [x_1\ x_2 \ldots x_p]'$ สามารถเขียนในรูป

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^p \mid \mathbf{\Sigma} \mid^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \tag{4}$$

โดยที่ $-\infty < x_i < \infty$, $i = 1, 2, ..., p$

อาจเขียนแทนด้วยสัญลักษณ์ $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$

สำหรับฟังก์ชันความน่าจะเป็นของการแจกแจงแบบปกติของ 2 ตัวแปร $(p = 2)$ กำหนดพารามิเตอร์ ดังนี้ $\mu_1 = E(x_1)$, $\mu_2 = E(x_2)$, $\sigma_{11} = \text{var}(x_1)$, $\sigma_{22} = \text{var}(x_2)$ และ $\rho_{12} = \sigma_{12}/(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}) = corr(x_1, x_2)$

เมตริกซ์ความแปรปรวนร่วม คือ

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

ซึ่งมีเมตริกซ์ผกผันของความแปรปรวนร่วม คือ

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

จาก $\rho_{12} = \sigma_{12}/(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}})$ จะได้ว่า $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$

ดังนั้น ระยะห่างกำลังสอง คือ

$$(\mathbf{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$=[x_1-\mu_1 \;\; x_2-\mu_2]\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}$$

$$=\frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$=\frac{1}{(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

พบว่า 2 เทอมสุดท้าย จะเป็นเทอมของค่ามาตรฐาน

เนื่องจาก $|\boldsymbol{\Sigma}|=\sigma_{11}\sigma_{22}-\sigma_{12}^2=\sigma_{11}\sigma_{22}(1-\rho_{12}^2)$ จะสามารถแทนค่า $\boldsymbol{\Sigma}^{-1}$ และ $|\boldsymbol{\Sigma}|$ ในสมการ (4) จะได้ฟังก์ชันความน่าจะเป็นของการแจกแจงแบบปกติของ 2 ตัวแปร ที่มีพารามิเตอร์ μ_1 , μ_2 , σ_{11} , σ_{22} และ ρ_{12} คือ

$$f(x_1,x_2)=\frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}$$
$$\times\exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

(Johnson and Wichern, 2007)

สำหรับสมบัติที่สำคัญบางประการของตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร (multivariate normal random variable) สามารถนำไปปรับใช้ในการพิสูจน์เพื่อสร้างตัวสถิติในการทดสอบภายใต้ข้อสมมติว่าข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร ดังนี้

กำหนดให้ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีขนาด $p\times 1$ มีการแจกแจงแบบปกติหลายตัวแปร อาจเขียนแทนด้วยสัญลักษณ์ $\mathbf{y}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ (Rencher, 2002)

1. ความเป็นปกติของผลบวกเชิงเส้นของตัวแปรสุ่ม $\mathbf{y}$ (normality of linear combinations of the variables in $\mathbf{y}$)

1.1 ถ้า $\mathbf{a}$ เป็นเวกเตอร์ของค่าคงที่ ฟังก์ชันเชิงเส้น $\mathbf{a}'\mathbf{y} = a_1y_1 + a_2y_2 + ... + a_py_p$ จะมีการแจกแจงแบบปกติตัวแปรเดียว (univariate normal distribution) อาจจะเขียนแทนด้วยสัญลักษณ์ $\mathbf{a}'\mathbf{y} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$

1.2 ถ้า $\mathbf{A}$ เป็นเมตริกซ์ของค่าคงที่ที่มีขนาด $q \times p$ เมื่อ $q \le p$ ผลบวกเชิงเส้นของ $\mathbf{Ay}$ จะมีการแจกแจงแบบปกติหลายตัวแปร อาจจะเขียนแทนด้วยสัญลักษณ์ $\mathbf{Ay} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

2. ตัวแปรมาตรฐาน (standardized variables)

เวกเตอร์ของตัวแปรมาตรฐาน $\mathbf{z}$ สามารถหาได้ 2 วิธี คือ

2.1 $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu})$ เมื่อ $\boldsymbol{\Sigma} = \mathbf{T}'\mathbf{T}$ ซึ่งแยกตัวประกอบด้วยวิธี Cholesky

2.2 $\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$ เมื่อ $\boldsymbol{\Sigma}^{1/2}$ คือ รากที่สองของเมตริกซ์ $\boldsymbol{\Sigma}$ (symmetric square root matrix of $\boldsymbol{\Sigma}$) หรือ $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$

เวกเตอร์ของตัวแปรมาตรฐาน $\mathbf{z}$ ที่ได้จากทั้งสองวิธี จะมีค่าเฉลี่ยเท่ากับ $\mathbf{0}$ ความแปรปรวนเท่ากับ $\mathbf{I}$ และความสัมพันธ์ระหว่างตัวแปรสุ่มทั้งหมดเท่ากับ 0 โดยที่ $\mathbf{z}$ จะมีการแจกแจงแบบปกติหลายตัวแปร อาจจะเขียนแทนด้วยสัญลักษณ์ $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$

3. การแจกแจงแบบไคกำลังสอง (Chi-square distribution)

ตัวแปรสุ่มไคกำลังสอง (Chi-square random variable) ที่มีองศาแห่งความอิสระ (degree of freedom) เท่ากับ p หาได้จากผลรวมกำลังสองของตัวแปรสุ่มปกติมาตรฐาน (standard normal random variables) ที่อิสระต่อกัน p ตัวแปร นั่นคือ ถ้า $\mathbf{z}$ เป็นเวกเตอร์ของตัวแปร

มาตรฐาน $\sum_{j=1}^{p} z_j^2 = \mathbf{z}'\mathbf{z}$ จะมีการแจกแจงแบบไคกำลังสองที่มีองศาแห่งความอิสระ p เขียนแทนด้วยสัญลักษณ์ χ_p^2 หรือ $\chi_{(p)}^2$

ดังนั้น ถ้า $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ และ $\mathbf{z}'\mathbf{z} = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ จะได้ว่า $(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi_p^2$

4. ความเป็นอิสระ (independence)

4.1 สำหรับเวกเตอร์ของตัวแปรสุ่ม $\mathbf{y}$ และ $\mathbf{x}$ ใดๆ ที่เป็นอิสระต่อกัน จะได้ว่าความแปรปรวนร่วมเท่ากับ $\mathbf{0}$ อาจจะเขียนแทนด้วยสัญลักษณ์ $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$

4.2 ตัวแปรสุ่ม y_j และ y_k เป็นอิสระต่อกัน เมื่อความแปรปรวนร่วมเท่ากับ 0 อาจจะเขียนแทนด้วยสัญลักษณ์ $\sigma_{jk} = 0$ โดยจะไม่เป็นจริงกรณีที่เป็นตัวแปรสุ่มไม่มีการแจกแจงแบบปกติ (nonnormal random variables)

5. การแจกแจงผลรวมของ 2 เวกเตอร์ย่อย (subvectors)

ถ้า $\mathbf{y}$ และ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปรที่มีขนาดเท่ากัน $(p \times 1)$ และเป็นอิสระต่อกัน จะได้ว่า

$$\mathbf{y} + \mathbf{x} \sim N_p(\boldsymbol{\mu}_y + \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx})$$

$$\mathbf{y} - \mathbf{x} \sim N_p(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx})$$

2. วิธีการวิเคราะห์องค์ประกอบหลัก

แนวคิดของวิธีการวิเคราะห์องค์ประกอบหลักคือ ต้องการลดขนาดของข้อมูลที่มีขนาดใหญ่ที่มีความสัมพันธ์ระหว่างกัน โดยทำการแปลงข้อมูลเป็นกลุ่มของตัวแปรใหม่ ที่เรียกว่าองค์ประกอบ (components) ซึ่งไม่มีความสัมพันธ์กัน

สมมติให้ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรสุ่ม p ตัวแปร โดยโครงสร้างของความแปรปรวนร่วมหรือความสัมพันธ์ระหว่างตัวแปรสุ่ม p ตัวแปร จะมีจำนวน $\frac{1}{2}p(p-1)$ คู่ วิธีการวิเคราะห์องค์ประกอบหลักจะมุ่งให้ความสนใจกับความแปรปรวนของตัวแปร

ขั้นตอนแรกของวิธีการวิเคราะห์องค์ประกอบหลักจะเริ่มต้นจากการหาฟังก์ชันเชิงเส้น (linear function) $\boldsymbol{\alpha}_1'\mathbf{x}$ ของ $\mathbf{x}$ ที่มีความแปรปรวนมากที่สุด โดยที่ $\boldsymbol{\alpha}_1$ เป็นเวกเตอร์ของ p ค่าคงที่ คือ $\alpha_{11},\alpha_{12},...,\alpha_{1p}$

$$\boldsymbol{\alpha}_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p = \sum_{j=1}^{p}\alpha_{1j}x_j$$

ขั้นตอนต่อไปจะทำการหาฟังก์ชันเชิงเส้น $\boldsymbol{\alpha}_2'\mathbf{x}$ ที่ไม่มีความสัมพันธ์กับ $\boldsymbol{\alpha}_1'\mathbf{x}$ และมีความแปรปรวนมากที่สุด และทำเช่นนี้ไปเรื่อยๆ จนกระทั่งขั้นตอนที่ i ก็จะได้ฟังก์ชันเชิงเส้น คือ $\boldsymbol{\alpha}_i'\mathbf{x}$ ที่มีความแปรปรวนมากที่สุด และไม่มีความสัมพันธ์กับ $\boldsymbol{\alpha}_1'\mathbf{x},\boldsymbol{\alpha}_2'\mathbf{x},...,\boldsymbol{\alpha}_{i-1}'\mathbf{x}$ โดยฟังก์ชันเหล่านี้จะเรียกว่า องค์ประกอบ ซึ่งถือว่าเป็นตัวแปรใหม่ และจะเรียก $\boldsymbol{\alpha}_i'\mathbf{x}$ ว่าองค์ประกอบหลักตัวที่ i

จากที่กล่าวมาข้างต้นสามารถอธิบายที่มาของการคำนวณได้ดังนี้

กำหนดให้ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่ทราบเมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}$ แต่โดยทั่วไปมักจะไม่ทราบค่า $\boldsymbol{\Sigma}$ ซึ่งสามารถประมาณด้วยเมตริกซ์ความแปรปรวนร่วมของตัวอย่าง ($\mathbf{S}$)

องค์ประกอบหลักตัวที่ i เขียนแทนด้วย $y_i = c$ เมื่อ $i = 1,2,...,p$ โดยที่ $\boldsymbol{\alpha}_i$ คือ เวกเตอร์เฉพาะของ $\boldsymbol{\Sigma}$ ที่สอดคล้องกับค่าเฉพาะ λ_k ที่มีค่าใหญ่ที่สุดในลำดับที่ i

ถ้ากำหนดให้ $\boldsymbol{\alpha}_i$ มีขนาดเท่ากับ 1 ($\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1$) แล้ว $\text{var}(y_i) = \lambda_i$ เมื่อ $\text{var}(y_i)$ แทนด้วยความแปรปรวนของ y_i

การพิสูจน์จะเริ่มจากการพิจารณา $\boldsymbol{\alpha}_1'\mathbf{x}$ โดย $\boldsymbol{\alpha}_1$ เป็นเวกเตอร์ที่มี $\text{var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ ค่ามากที่สุด ซึ่งจะสามารถมีคำตอบของ $\boldsymbol{\alpha}_1$ ได้จำนวนไม่จำกัด จึงจะทำการปรับเงื่อนไขให้เป็น

บรรทัดฐาน (normalization constraint) หรือ $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ ซึ่งในการกำหนดเงื่อนไขในลักษณะอื่นอาจจะทำให้การคำนวณมีความยุ่งยากมากขึ้น

การหาค่ามากที่สุดของ $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ เมื่อกำหนดเงื่อนไข $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ วิธีการแก้ปัญหาที่นำมาใช้คือ ตัวคูณลากรานจ์ (Lagrange multiplier)

$$Maximize[\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1)] \tag{5}$$

เมื่อ λ คือ ตัวคูณลากรานจ์

จากสมการ (5) เมื่อหาอนุพันธ์เทียบกับ $\boldsymbol{\alpha}_1$ จะได้

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_1 = 0 \tag{6}$$

เมื่อ $\mathbf{I}_p$ คือ เมตริกซ์เอกลักษณ์ที่มีขนาด $p \times p$

λ คือ ค่าเฉพาะของ $\boldsymbol{\Sigma}$ ที่มีเวกเตอร์เฉพาะ $\boldsymbol{\alpha}_1$

จากสมการ (6) เมื่อจัดรูปสมการใหม่ และคูณ $\boldsymbol{\alpha}_1'$ ทางซ้ายของสมการ จะได้ว่า

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda \tag{7}$$

ดังนั้น λ จะมีค่ามากเท่าที่จะเป็นไปได้ ด้วยเหตุนี้ $\boldsymbol{\alpha}_1$ จะเป็นเวกเตอร์เฉพาะที่สอดคล้องกับค่าเฉพาะที่มากที่สุดของ $\boldsymbol{\Sigma}$ และ $\text{var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 = \lambda_1$ เป็นค่าเฉพาะที่มากที่สุด

สำหรับองค์ประกอบหลักตัวที่ 2 คือการหาค่ามากที่สุดของ $\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ ที่ไม่มีความสัมพันธ์กับ $\boldsymbol{\alpha}_1'\mathbf{x}$ หรือ $\text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) = 0$ โดยที่ $\text{cov}(x, y)$ แทนความแปรปรวนร่วมระหว่างตัวแปรสุ่ม x และ y จะได้ว่า

$$\text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda_1\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 \tag{8}$$

จากสมการ (8) และ $\text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) = 0$ จะได้ว่า

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = 0 \tag{9}$$

การหาค่ามากที่สุดของ $\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ ภายใต้กำหนดเงื่อนไข $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1$ และ $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = 0$ จากสมการ (9) จะได้ว่า

$$Maximize[\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \varphi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1] \tag{10}$$

เมื่อ λ และ φ คือ ตัวคูณลากรานจ์

จากสมการ (10) เมื่อหาอนุพันธ์เทียบกับ $\boldsymbol{\alpha}_2$ ทำการจัดรูปสมการใหม่ และคูณ $\boldsymbol{\alpha}_1'$ ทางซ้ายของสมการ จะได้ว่า

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 - \varphi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0 \tag{11}$$

แต่เนื่องจากสมการ (9) คือ $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0$ และ $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ จึงทำให้ $\varphi = 0$

ดังนั้น จากอนุพันธ์เทียบกับ $\boldsymbol{\alpha}_2$ ของสมการ (10) จะได้ว่า

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_2 = 0 \tag{12}$$

จากสมการ (12) เมื่อจัดรูปสมการใหม่ และคูณ $\boldsymbol{\alpha}_2'$ ทางซ้ายของสมการ จะได้ว่า

$$\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\lambda\boldsymbol{\alpha}_2 = \lambda\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = \lambda \tag{13}$$

ดังนั้น λ จะมีค่ามากเท่าที่จะเป็นไปได้ โดยสมมติว่าไม่มีค่าเฉพาะที่ซ้ำกัน $\lambda \neq \lambda_1$ เนื่องจากกรณีที่มีค่าเฉพาะซ้ำกัน วิธีการคำนวณจะมีความยุ่งยากมากขึ้น และจะมีผลกระทบต่อเงื่อนไข $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0$ ทำให้ไม่เป็นจริง นั่นคือ λ จะเป็นค่าเฉพาะที่มีค่ามากเป็นลำดับที่สอง และมี $\boldsymbol{\alpha}_2$ เป็นเวกเตอร์เฉพาะที่สอดคล้องกัน

สำหรับองค์ประกอบหลักตัวที่ 3 , 4 จนถึงตัวที่ p จะมี $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, ..., \boldsymbol{\alpha}_p$ เป็นเวกเตอร์เฉพาะของ $\boldsymbol{\Sigma}$ ที่สอดคล้องกับ $\lambda_3, \lambda_4, ..., \lambda_p$ ซึ่งเป็นค่าเฉพาะตัวที่ 3, 4, …, p ซึ่งเป็นค่าเฉพาะเรียงค่าจากมากไปหาน้อยที่สุด และจะได้ว่า $\text{var}(y_i) = \lambda_i$ สำหรับ $i = 1, 2, ..., p$ (Jolliffe, 2002)

การวิเคราะห์องค์ประกอบหลักเป็นการวิเคราะห์เกี่ยวกับการอธิบายโครงสร้างของความแปรปรวนของตัวแปรทั้งหมดด้วยฟังก์ชันเชิงเส้นของตัวแปรเดิมเพียง 2-3 ฟังก์ชัน โดยจะต้องมีรายละเอียดหรือความแปรปรวนจากตัวแปรเดิมมาไว้ในตัวแปรใหม่ให้มากที่สุด และไม่จำเป็นต้องทำภายใต้ข้อสมมติว่าข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร

สามารถเขียนองค์ประกอบให้อยู่ในรูปทั่วไปได้ คือ

$$\begin{aligned} y_1 &= \boldsymbol{\alpha}_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p \\ y_2 &= \boldsymbol{\alpha}_2'\mathbf{x} = \alpha_{21}x_1 + \alpha_{22}x_2 + ... + \alpha_{2p}x_p \\ &\vdots \\ y_p &= \boldsymbol{\alpha}_p'\mathbf{x} = \alpha_{p1}x_1 + \alpha_{p2}x_2 + ... + \alpha_{pp}x_p \end{aligned}$$

และ $\text{var}(y_i) = \boldsymbol{\alpha}_i'\Sigma\boldsymbol{\alpha}_i \qquad i = 1, 2, ..., p$

$\text{cov}(y_i, y_k) = \boldsymbol{\alpha}_i'\Sigma\boldsymbol{\alpha}_k \qquad i \neq k$

องค์ประกอบที่ 1 เป็นฟังก์ชันเชิงเส้นของตัวแปรเดิม ซึ่งให้ค่าความแปรปรวนมากที่สุด และกำหนดให้ขนาดของเวกเตอร์เฉพาะมีค่าเท่ากับ 1 จะสามารถเขียนแทนได้ว่า องค์ประกอบที่ 1 คือ ฟังก์ชันเชิงเส้น $\boldsymbol{\alpha}_1'\mathbf{x}$ ที่ทำให้ $\text{var}(\boldsymbol{\alpha}_1'\mathbf{x})$ มีค่ามากที่สุด ซึ่งมี $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$

สำหรับองค์ประกอบที่ 2 คือ ฟังก์ชันเชิงเส้น $\boldsymbol{\alpha}_2'\mathbf{x}$ ที่ทำให้ $\text{var}(\boldsymbol{\alpha}_2'\mathbf{x})$ มีค่ามากที่สุด ซึ่งมี $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1$ และ $\text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) = 0$ และองค์ประกอบอื่นๆ ก็จะมีลักษณะเช่นเดียวกัน

การพิจารณาจำนวนตัวแปรใหม่หรือองค์ประกอบหลักที่เหมาะสม สามารถพิจารณาจากร้อยละความแปรปรวนสะสม โดยจะเลือกจำนวนองค์ประกอบหลักที่มีร้อยละความแปรปรวนสะสมอย่างน้อยร้อยละ 80 อีกแนวทางหนึ่งคือเลือกองค์ประกอบหลักที่มีค่าเฉพาะมากกว่าค่าเฉลี่ยของค่าเฉพาะทั้งหมด หรืออาจจะพิจารณาจาก scree plot และจากวิธีการทดสอบสมมติฐาน ภายใต้ข้อสมมติว่าข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร (Rencher, 1998)

3. วิธีที่นำเสนอโดย Hawkins

Hawkins (1974) เสนอแนวคิดของวิธีการวิเคราะห์องค์ประกอบหลัก และพัฒนาตัวสถิติมาจากวิธีที่เสนอโดย Rao (1964) ที่ได้สร้างตัวสถิติ d_{1i}^2 ซึ่งเป็นตัวสถิติที่ให้ความสำคัญกับองค์ประกอบรองมากกว่าองค์ประกอบหลัก จากนั้น Gnanadesikan and Kettenring (1972) ได้ทำการศึกษาตัวสถิติ d_{1i}^2 เพิ่มเติม และกำหนดให้อยู่ในรูปของค่าองค์ประกอบกำลังสอง

$$d_{1i}^2 = \sum_{k=p-q+1}^{p} z_{ik}^2$$

เมื่อ z_{ik} คือ ค่าองค์ประกอบที่ k สำหรับข้อมูลชุดที่ i

q คือ จำนวนองค์ประกอบรองที่ใช้ในการตรวจสอบค่าผิดปกติ

p คือ จำนวนองค์ประกอบทั้งหมด

สำหรับรูปแบบการแจกแจงของตัวสถิติ d_{1i}^2 นี้สามารถประมาณการแจกแจงได้เป็นการแจกแจงแบบแกมมา ดังนั้นตัวสถิติ d_{1i}^2 จะมีการแจกแจงแบบแกมมา ถ้าข้อมูลไม่มีค่าผิดปกติ โดยจะทำการพิจารณาจาก gamma probability plot เพื่อใช้ตรวจสอบค่าผิดปกติ อย่างไรก็ตาม ในงานวิจัยนี้ไม่มีการเสนอวิธีการทดสอบที่เป็นทางการ (formal approach) แต่เสนอวิธีการตรวจสอบโดยใช้กราฟ ด้วยการลงจุดรายคู่ขององค์ประกอบที่อยู่ติดกันซึ่งทำได้ง่าย ซึ่งการตรวจสอบจากกราฟยังให้ผลไม่ดีเท่าที่ควร (Hawkins, 1980)

ในปี ค.ศ. 1974 Hawkins เสนอตัวสถิติ d_{2i}^2 เพื่อแก้ปัญหาความสำคัญขององค์ประกอบแตกต่างกันเมื่อนำมาใช้กับตัวสถิติ d_{1i}^2 โดยเฉพาะยิ่งเมื่อใช้จำนวนองค์ประกอบรองมากจนเข้าใกล้องค์ประกอบทั้งหมด ($q \to p$) เนื่องจากองค์ประกอบรองจะมีความแปรปรวนลดลงไปเรื่อยๆ จึงทำให้เกิดแนวคิดในการถ่วงน้ำหนักให้แต่ละองค์ประกอบมีน้ำหนักที่แตกต่างกัน โดยถ่วงน้ำหนักจากค่าเฉพาะขององค์ประกอบนั้นๆ

$$d_{2i}^2 = \sum_{k=p-q+1}^{p} \frac{z_{ik}^2}{l_k}$$

เมื่อ l_k คือ ค่าเฉพาะขององค์ประกอบที่ k

จากการศึกษารูปแบบการแจกแจงของตัวสถิติ d_{2i}^2 พบว่า มีการแจกแจงแบบไคกำลังสองที่มีองศาแห่งความอิสระเท่ากับจำนวนองค์ประกอบรองที่ใช้ (q) อาจเขียนแทนด้วยสัญลักษณ์ $d_{2i}^2 \sim \chi_q^2$

ดังนั้น เมื่อใช้องค์ประกอบรอง 2 องค์ประกอบ จะได้ว่า

$$d_{2i}^2 \sim \chi_2^2$$

และเมื่อใช้องค์ประกอบรอง 3 องค์ประกอบ จะได้ว่า

$$d_{2i}^2 \sim \chi_3^2$$

อย่างไรก็ตาม ปัจจุบันยังคงมีปัญหาเรื่องจำนวนองค์ประกอบรองที่เหมาะสมกับตัวสถิติดังกล่าว ซึ่งเป็นปัญหาที่ยากมาก ทั้งยังไม่มีผู้ใดสามารถหาหลักการที่ชัดเจนได้ Jolliffe (2002) กล่าวว่าจำนวนองค์ประกอบรอง 2-3 องค์ประกอบสุดท้าย จะสามารถใช้ตรวจค่าผิดปกติในลักษณะที่ความสัมพันธ์ของตัวแปรไม่สอดคล้องกันได้ดี

สำหรับตัวสถิติ d_{2i}^2 ที่ใช้ในการเปรียบเทียบในงานวิจัยนี้จะใช้ 2 กรณี ได้แก่ เมื่อใช้องค์ประกอบรอง 2 และ 3 องค์ประกอบ ดังนั้นในการศึกษาวิจัยครั้งนี้จะทำการกำหนดสัญลักษณ์ของตัวสถิติดังกล่าวขึ้นใหม่ เพื่อให้เกิดความแตกต่างของตัวสถิติทั้ง 2 กรณีอย่างชัดเจน และไม่ให้เกิดความสับสน

นั่นคือ สัญลักษณ์ของตัวสถิติ d_{2i}^2 เมื่อใช้องค์ประกอบรอง 2 องค์ประกอบ คือ

$$d_{2i,(2)}^2 \sim \chi_2^2$$

และสัญลักษณ์ของตัวสถิติ d_{2i}^2 เมื่อใช้องค์ประกอบรอง 3 องค์ประกอบ คือ

$$d_{2i,(3)}^2 \sim \chi_3^2$$

4. วิธี Mahalanobis distance

วิธีนี้ถูกเสนอในปี ค.ศ. 1936 โดย Prasanta Chandra Mahalanobis นักสถิติชาวอินเดีย (Mahalanobis, 1936) สามารถนำมาใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยการวัดระยะห่างระหว่างข้อมูลกับค่าเฉลี่ยของข้อมูล (Jobson, 1992) และจะให้ความสำคัญกับความแปรปรวนร่วม ซึ่งแตกต่างจาก Euclidean distance ที่หาได้จาก

$$\text{Euclidean distance} = (\mathbf{y}-\boldsymbol{\mu})'(\mathbf{y}-\boldsymbol{\mu})$$

สำหรับ Mahalanobis distance (d^2) จะทำการปรับให้ระยะห่างเป็นค่ามาตรฐานด้วยเมตริกซ์ผกผันของความแปรปรวนร่วม คือ

$$d^2 = (\mathbf{y}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

กรณีที่ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร พบว่า d^2 จะมีการแจกแจงแบบไคกำลังสอง ที่มีองศาแห่งความอิสระเท่ากับจำนวนตัวแปร (p) อาจเขียนแทนด้วยสัญลักษณ์ $d^2 \sim \chi_p^2$

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องไมโครคอมพิวเตอร์ ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

2. โปรแกรม SAS version 9.1

วิธีการ

1. ศึกษาหลักการ และทฤษฎีที่เกี่ยวข้องกับวิธีการวิเคราะห์องค์ประกอบหลัก เพื่อสร้างตัวสถิติที่นำองค์ประกอบรอง 2 และ 3 องค์ประกอบ มาใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร ภายใต้ข้อสมมติว่าข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร

สำหรับขั้นตอนการสร้างตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร กรณีเมื่อใช้องค์ประกอบรอง 2 องค์ประกอบ ทำโดยการนำความรู้ทางทฤษฎีและสมบัติของการแจกแจงแบบปกติมาปรับใช้ สามารถเขียนเป็นทฤษฎีบทได้ดังนี้

ทฤษฎีบท 1 กำหนดให้ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีขนาด $p\times 1$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}$ และเมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}$ เขียนแทนด้วยสัญลักษณ์ $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ เมื่อทำการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ จะมีคู่อันดับของค่าเฉพาะและเวกเตอร์เฉพาะ $(\lambda_1, \boldsymbol{\alpha}_1), (\lambda_2, \boldsymbol{\alpha}_2), ..., (\lambda_p, \boldsymbol{\alpha}_p)$ เมื่อ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ จะได้องค์ประกอบทั้งหมด p องค์ประกอบ นั่นคือ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีขนาด $p\times 1$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีเวกเตอร์ค่าเฉลี่ย $\mathbf{0}$ และเมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}_y$ เขียนแทนด้วยสัญลักษณ์ $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_y)$

โดยที่ $\boldsymbol{\Sigma}_y = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$ ซึ่ง $y_{p-1} \sim N(0, \lambda_{p-1})$ และ $y_p \sim N(0, \lambda_p)$

จะได้ตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร เมื่อใช้องค์ประกอบรอง 2 องค์ประกอบ คือ $R^2_{2z} = \dfrac{R_2^2}{\lambda_{p-1} + \lambda_p} \sim \chi^2_{(1)}$

พิสูจน์

กำหนดให้ $\mathbf{x}=[x_1\ x_2\ \ldots\ x_p]'$ มีการแจกแจงแบบปกติหลายตัวแปร ด้วยค่าเฉลี่ย $\boldsymbol{\mu}=[\mu_1\ \mu_2\ \ldots\ \mu_p]'$ และความแปรปรวน $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}$ อาจเขียนแทนด้วย สัญลักษณ์ $\mathbf{x}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$

ฟังก์ชันการแจกแจงความน่าจะเป็น คือ

$$f(\mathbf{x})=\frac{1}{(\sqrt{2\pi})^p\,|\boldsymbol{\Sigma}|^{1/2}}e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

โดยที่ p แทนจำนวนตัวแปร

เมื่อ $\mathbf{x}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ สามารถเขียนเวกเตอร์มาตรฐาน (standardized vector) ได้ดังนี้

$$\mathbf{z}=(\mathbf{V}^{\frac{1}{2}})^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

โดยที่ $\mathbf{V}^{\frac{1}{2}}=\begin{bmatrix}\sqrt{\sigma_{11}} & 0 & \cdots & 0\\ 0 & \sqrt{\sigma_{22}} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}}\end{bmatrix}$

จะได้ว่า $E(\mathbf{z})=\mathbf{0}$ และ $\operatorname{cov}(\mathbf{z})=(\mathbf{V}^{\frac{1}{2}})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{\frac{1}{2}})=\boldsymbol{\rho}$

โดยที่ $\boldsymbol{\rho}$ แทนเมตริกซ์สหสัมพันธ์

อาจเขียนแทนด้วยสัญลักษณ์ $\mathbf{z}\sim N_p(\mathbf{0},\boldsymbol{\rho})$

กำหนดให้ $\boldsymbol{\rho}$ เป็นเมตริกซ์ความแปรปรวนร่วมของ $\mathbf{z} = [z_1\ z_2\ \ldots\ z_p]'$ ที่มีคู่อันดับของค่าเฉพาะและเวกเตอร์เฉพาะ $(\lambda_1, \boldsymbol{\alpha}_1), (\lambda_2, \boldsymbol{\alpha}_2), \ldots, (\lambda_p, \boldsymbol{\alpha}_p)$ เมื่อ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$

สามารถเขียนองค์ประกอบที่ i ให้อยู่ในรูปทั่วไป คือ

$$y_i = \boldsymbol{\alpha}_i'\mathbf{z} = \boldsymbol{\alpha}_{i1}z_1 + \boldsymbol{\alpha}_{i2}z_2 + \ldots + \boldsymbol{\alpha}_{ip}z_p$$

เมื่อ $i = 1, 2, \ldots, p$

จะได้ว่า

$$\begin{aligned} \text{var}(y_i) &= \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_i = \lambda_i && i = 1, 2, \ldots, p \\ \text{cov}(y_i, y_k) &= \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_k = 0 && i \neq k \end{aligned}$$

ถ้าในกรณีที่ λ_i มีค่าซ้ำกัน จะทำให้ค่า y_i มีค่าได้มากกว่าหนึ่งค่า แต่ถ้า $\lambda_1, \lambda_2, \ldots, \lambda_p$ มีค่าแตกต่างกัน ค่าเวกเตอร์เฉพาะที่ได้จาก $\boldsymbol{\rho}$ จะตั้งฉากกัน (orthogonal) และในทางปฏิบัติพบว่าค่าเฉพาะของเมตริกซ์สหสัมพันธ์จะมีค่าแตกต่างกัน และจะไม่เป็นศูนย์ (Jolliffe, 2002)

นั่นคือ สำหรับเวกเตอร์ค่าเฉพาะ $\boldsymbol{\alpha}_i$ และ $\boldsymbol{\alpha}_k$ ใดๆ จะได้ว่า $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_k = 0,\ i \neq k$

จาก $\boldsymbol{\rho}\boldsymbol{\alpha}_k = \lambda_k\boldsymbol{\alpha}_k$ เมื่อนำ $\boldsymbol{\alpha}_i'$ คูณทางซ้ายตลอดสมการ จะได้

$$\text{cov}(y_i, y_k) = \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_i'\lambda_k\boldsymbol{\alpha}_k = 0,\ i \neq k$$

พิจารณาผลบวกเชิงเส้น

$$\begin{aligned} y_1 &= \boldsymbol{\alpha}_1'\mathbf{z} = \alpha_{11}z_1 + \alpha_{12}z_2 + \ldots + \alpha_{1p}z_p \\ y_2 &= \boldsymbol{\alpha}_2'\mathbf{z} = \alpha_{21}z_1 + \alpha_{22}z_2 + \ldots + \alpha_{2p}z_p \\ &\vdots \qquad\qquad\qquad \vdots \\ y_p &= \boldsymbol{\alpha}_p'\mathbf{z} = \alpha_{p1}z_1 + \alpha_{p2}z_2 + \ldots + \alpha_{pp}z_p \end{aligned}$$

สามารถเขียนผลบวกเชิงเส้นให้อยู่ในรูปเมตริกซ์ ดังนี้

$$\mathbf{y} = \boldsymbol{\alpha}\mathbf{z}$$

เมื่อ $\mathbf{y} = [y_1\ y_2 \ldots y_p]'$ และ $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\rho})$

$\mathbf{z} = [z_1\ z_2 \ldots z_p]'$

พิจารณา

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \ldots & \alpha_{pp} \end{bmatrix}$$

เมื่อ $\boldsymbol{\alpha}$ คือ เมตริกซ์ของค่าคงที่มีขนาด $p \times p$ และมีค่าลำดับชั้น (rank) เท่ากับ p (Johnson and Wichern, 2007)

ดังนั้น ผลบวกเชิงเส้นในรูป $\boldsymbol{\alpha}\mathbf{z}$ จำนวน p สมการ มีการแจกแจงแบบปกติหลายตัวแปร ซึ่งจากสมบัติของตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร จะได้ว่า ถ้า $\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\rho})$ แล้ว $\boldsymbol{\alpha}\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\alpha\rho\alpha}')$ นั่นคือ $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\alpha\rho\alpha}')$ (Rencher, 2002)

เนื่องจาก $\operatorname{var}(y_i) = \lambda_i \qquad i = 1, 2, \ldots, p$

$\operatorname{cov}(y_i, y_k) = 0 \qquad i \neq k$

จะได้

$$\boldsymbol{\alpha\rho\alpha}' = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}$$

ซึ่งเป็นเมตริกซ์ความแปรปรวนร่วมของ $\mathbf{y}$ เขียนแทนด้วย $\boldsymbol{\Sigma}_y$

ถ้า y_i และ y_k มีการแจกแจงแบบปกติแบบสองตัวแปร (bivariate normal distribution) และ $\text{cov}(y_i, y_k) = 0$ แล้ว y_i และ y_k จะเป็นอิสระต่อกัน

ถ้า $\mathbf{y}$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีค่าเฉลี่ยเป็น $\mathbf{0}$ และความแปรปรวนร่วมเป็นเมตริกซ์ย่อยของ $\mathbf{\Sigma}_y$ ที่สอดคล้องกัน เซตย่อยใดๆ ของ $\mathbf{y}$ มีการแจกแจงแบบปกติ แต่ไม่เป็นจริงในทางกลับกัน กล่าวคือ y ใดๆ ที่มีการแจกแจงแบบปกติแล้ว ไม่จำเป็นจะต้องมี $\mathbf{y}$ ที่มีการแจกแจงแบบปกติหลายตัวแปร นั่นคือ ถ้า $\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{\Sigma}_y)$ แต่ละ y_i ใน $\mathbf{y}$ จะมีการแจกแจงแบบปกติ นั่นคือ $y_i \sim N(0, \lambda_i)$ เมื่อ $i = 1, 2, ..., p$ (Rencher, 2002)

พิจารณาองค์ประกอบรอง 2 องค์ประกอบ นั่นคือ y_{p-1} และ y_p

จาก $y_i \sim N(0, \lambda_i),\ i = 1, 2, \ldots, p$

จะได้

$$y_{p-1} \sim N(0, \lambda_{p-1})$$

และ

$$y_p \sim N(0, \lambda_p)$$

ถ้า y_{p-1} และ y_p มีขนาดเท่ากัน และเป็นอิสระต่อกัน แล้ว $(y_{p-1} + y_p) \sim N(0, \lambda_{p-1} + \lambda_p)$ (Rencher, 2002)

กำหนดให้ $R_2 = y_{p-1} + y_p$

ดังนั้น

$$R_2 \sim N(0, \lambda_{p-1} + \lambda_p)$$

$$\frac{R_2}{\sqrt{\lambda_{p-1} + \lambda_p}} \sim N(0,1)$$

นั่นคือ

$$R_{2z}^2 = \frac{R_2^2}{\lambda_{p-1} + \lambda_p} \sim \chi_{(1)}^2$$

ทฤษฎีบท 2 กำหนดให้ $\mathbf{x}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีขนาด $p \times 1$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}$ และเมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}$ เขียนแทนด้วยสัญลักษณ์ $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ เมื่อทำการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ จะมีคู่อันดับของค่าเฉพาะและเวกเตอร์เฉพาะ $(\lambda_1, \boldsymbol{\alpha}_1), (\lambda_2, \boldsymbol{\alpha}_2), ..., (\lambda_p, \boldsymbol{\alpha}_p)$ เมื่อ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ จะได้องค์ประกอบทั้งหมด p องค์ประกอบ นั่นคือ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีขนาด $p \times 1$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีเวกเตอร์ค่าเฉลี่ย $\mathbf{0}$ และเมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}_y$ เขียนแทนด้วยสัญลักษณ์ $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_y)$

โดยที่ $\boldsymbol{\Sigma}_y = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$ ซึ่ง $y_{p-2} \sim N(0, \lambda_{p-2}), y_{p-1} \sim N(0, \lambda_{p-1})$ และ $y_p \sim N(0, \lambda_p)$

จะได้ตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร เมื่อใช้องค์ประกอบรอง 3 องค์ประกอบ คือ $R_{3z}^2 = \frac{R_3^2}{\lambda_{p-2} + \lambda_{p-1} + \lambda_p} \sim \chi_{(1)}^2$

พิสูจน์

สำหรับตัวสถิติที่ใช้องค์ประกอบรอง 3 องค์ประกอบ สามารถพิสูจน์ได้ในทำนองเดียวกันกับตัวสถิติที่ใช้องค์ประกอบรอง 2 องค์ประกอบ

จาก $y_{p-2} \sim N(0, \lambda_{p-2}), y_{p-1} \sim N(0, \lambda_{p-1})$ และ $y_p \sim N(0, \lambda_p)$

กำหนดให้ $R_3 = y_{p-2} + y_{p-1} + y_p$

ดังนั้น $R_3 \sim N(0, \lambda_{p-2} + \lambda_{p-1} + \lambda_p)$

$$\frac{R_3}{\sqrt{\lambda_{p-2}+\lambda_{p-1}+\lambda_p}} \sim N(0,1)$$

นั่นคือ

$$R_{3z}^2 = \frac{R_3^2}{\lambda_{p-2}+\lambda_{p-1}+\lambda_p} \sim \chi_{(1)}^2$$

ตัวสถิติ R_{2z}^2 และ R_{3z}^2 เป็นตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยใช้องค์ประกอบรอง 2 และ 3 องค์ประกอบ ตามลำดับ โดยสามารถตรวจสอบค่าผิดปกติที่ไม่ปรากฏชัดเมื่อพิจารณาความผิดปกติของข้อมูลในแต่ละตัวแปร แต่เป็นค่าผิดปกติในลักษณะที่ความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน

2. จำลองข้อมูลโดยจำลองสถานการณ์ กำหนดค่าต่างๆ ดังนี้

2.1 ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร

การสร้างตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร จะต้องกำหนด $\boldsymbol{\mu}$ และ $\boldsymbol{\Sigma}$ ซึ่งเป็นพารามิเตอร์ที่ใช้ในการแจกแจง จากนั้นจะใช้หลักการดังนี้

กำหนดให้ $\mathbf{y} = [y_1\ y_2 \ldots y_p]'$ มีการแจกแจงแบบปกติหลายตัวแปร ด้วยค่าเฉลี่ย $\boldsymbol{\mu} = [\mu_1\ \mu_2 \ldots \mu_p]'$ และความแปรปรวน $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$ โดย $y_i \sim N(\mu_i, \sigma_{ii})$

ที่มี $\mathrm{cov}(y_i, y_j) = \sigma_{ij}$ และ $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$

สร้างเวกเตอร์สุ่ม $\mathbf{z} = [z_1\ z_2 \ldots z_p]'$ โดย $z_i \sim N(0,1)$ และเป็นอิสระต่อกัน จะได้ว่า $\mathrm{cov}(z_i, z_j) = 0$ ซึ่งเมตริกซ์ความแปรปรวนร่วมจะเป็นเมตริกซ์เอกลักษณ์ เขียนแทนสัญลักษณ์ $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$

เมื่อกำหนด $\boldsymbol{\Sigma}$ เป็นเมตริกซ์ความแปรปรวนร่วม จะเป็นเมตริกซ์สมมาตรที่มีขนาด $p \times p$ และเป็นเมตริกซ์บวกแน่นอน (positive-definite matrix) จะมี $\mathbf{T}$ เป็นเมตริกซ์สามเหลี่ยมบน (upper-triangular matrix) ที่มีขนาด $p \times p$ ที่ทำให้ $\mathbf{T}'\mathbf{T} = \boldsymbol{\Sigma}$

หลังจากที่ได้ $\mathbf{T}$ แล้วจะทำการแปลงเวกเตอร์สุ่มที่มีการแจกแจงแบบปกติมาตรฐานให้มีการแจกแจงแบบปกติหลายตัวแปรที่มีพารามิเตอร์ตามที่กำหนด

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{T}'\mathbf{z}$$

จะได้ว่า $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

สำหรับกรณีที่ต้องการสร้างตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร เมื่อกำหนดเมตริกซ์สหสัมพันธ์

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

จะสามารถนำเมตริกซ์สหสัมพันธ์ไปคำนวณร่วมกับความแปรปรวนของแต่ละตัวแปร เพื่อหาเมตริกซ์ความแปรปรวนร่วมได้ไม่ยาก เพื่อนำไปใช้ในการกำหนดเป็นพารามิเตอร์ในการสร้างตัวแปรสุ่ม โดยที่

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{pp} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{pp} \end{bmatrix}$$

2.2 กำหนดจำนวนตัวแปร $p =$ 5, 10(2)20

2.3 ขนาดข้อมูลตัวอย่าง $n =$ 30, 40, 50, 75, 100

2.4 ร้อยละของค่าผิดปกติที่เกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน 4 ระดับ ได้แก่ 10, 20, 30 และ 40

ข้อมูลที่มีค่าผิดปกติจะมาจากประชากรที่มีการแจกแจงแบบปกติหลายตัวแปรที่มีค่าเฉลี่ย และความแปรปรวนเช่นเดียวกับข้อมูลปกติ แต่จะมีความแปรปรวนร่วมแตกต่างกัน ซึ่งจะส่งผลทำให้รูปแบบความสัมพันธ์ของตัวแปรแตกต่างไปจากข้อมูลปกติ โดยค่าพารามิเตอร์ในแต่ละสถานการณ์ที่กำหนดขึ้นในงานวิจัยนี้สร้างขึ้นโดยการสุ่มด้วยโปรแกรมคอมพิวเตอร์

เช่น กรณี $p = 10,\ n = 30$ และร้อยละของค่าผิดปกติเท่ากับ 10 เป็นข้อมูลที่มาจากประชากรที่มีการแจกแจงแบบปกติหลายตัวแปร 10 ตัวแปร ประกอบด้วย 30 ชุดข้อมูล โดยข้อมูลแบ่งออกเป็น 2 กลุ่ม กลุ่มแรกได้แก่ ข้อมูล 27 ชุด เป็นข้อมูลปกติ และกลุ่มที่สองได้แก่ ข้อมูล 3 ชุด หรือ 10% ของข้อมูลทั้งหมด จะเป็นข้อมูลที่ผิดปกติ สามารถเขียนได้ดังนี้

กลุ่มแรกเป็นข้อมูลปกติ ประกอบด้วยข้อมูล 27 ชุด นั่นคือ $\mathbf{y} = [y_1\ y_2 \ldots y_{10}]'$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีค่าเฉลี่ย $\boldsymbol{\mu} = [\mu_1\ \mu_2 \ldots \mu_{10}]'$ และความแปรปรวน

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1,10} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{10,1} & \sigma_{10,2} & \cdots & \sigma_{10,10} \end{bmatrix}$$ สามารถเขียนแทนด้วยสัญลักษณ์ $\mathbf{y} \sim N_{10}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

มีรูปแบบความสัมพันธ์คือ $\boldsymbol{\rho} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1,10} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{10,1} & \rho_{10,2} & \cdots & \rho_{10,10} \end{bmatrix}$

สำหรับกลุ่มที่สองจะเป็นข้อมูลผิดปกติ ซึ่งประกอบด้วยข้อมูล 3 ชุด นั่นคือ $\mathbf{y} = [y_1\ y_2 \ldots y_{10}]'$ มีการแจกแจงแบบปกติหลายตัวแปร ที่มีค่าเฉลี่ย $\boldsymbol{\mu} = [\mu_1\ \mu_2 \ldots \mu_{10}]'$ และความแปรปรวน $\boldsymbol{\Sigma}' = \begin{bmatrix} \sigma_{11} & \sigma'_{12} & \cdots & \sigma'_{1,10} \\ \sigma'_{12} & \sigma_{22} & \cdots & \sigma'_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma'_{10,1} & \sigma'_{10,2} & \cdots & \sigma_{10,10} \end{bmatrix}$ สามารถเขียนแทนด้วยสัญลักษณ์ $\mathbf{y} \sim N_{10}(\boldsymbol{\mu}, \boldsymbol{\Sigma}')$

มีรูปแบบความสัมพันธ์คือ $\boldsymbol{\rho}' = \begin{bmatrix} \rho_{11} & \rho'_{12} & \cdots & \rho'_{1,10} \\ \rho'_{12} & \rho_{22} & \cdots & \rho'_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ \rho'_{10,1} & \rho'_{10,2} & \cdots & \rho_{10,10} \end{bmatrix}$

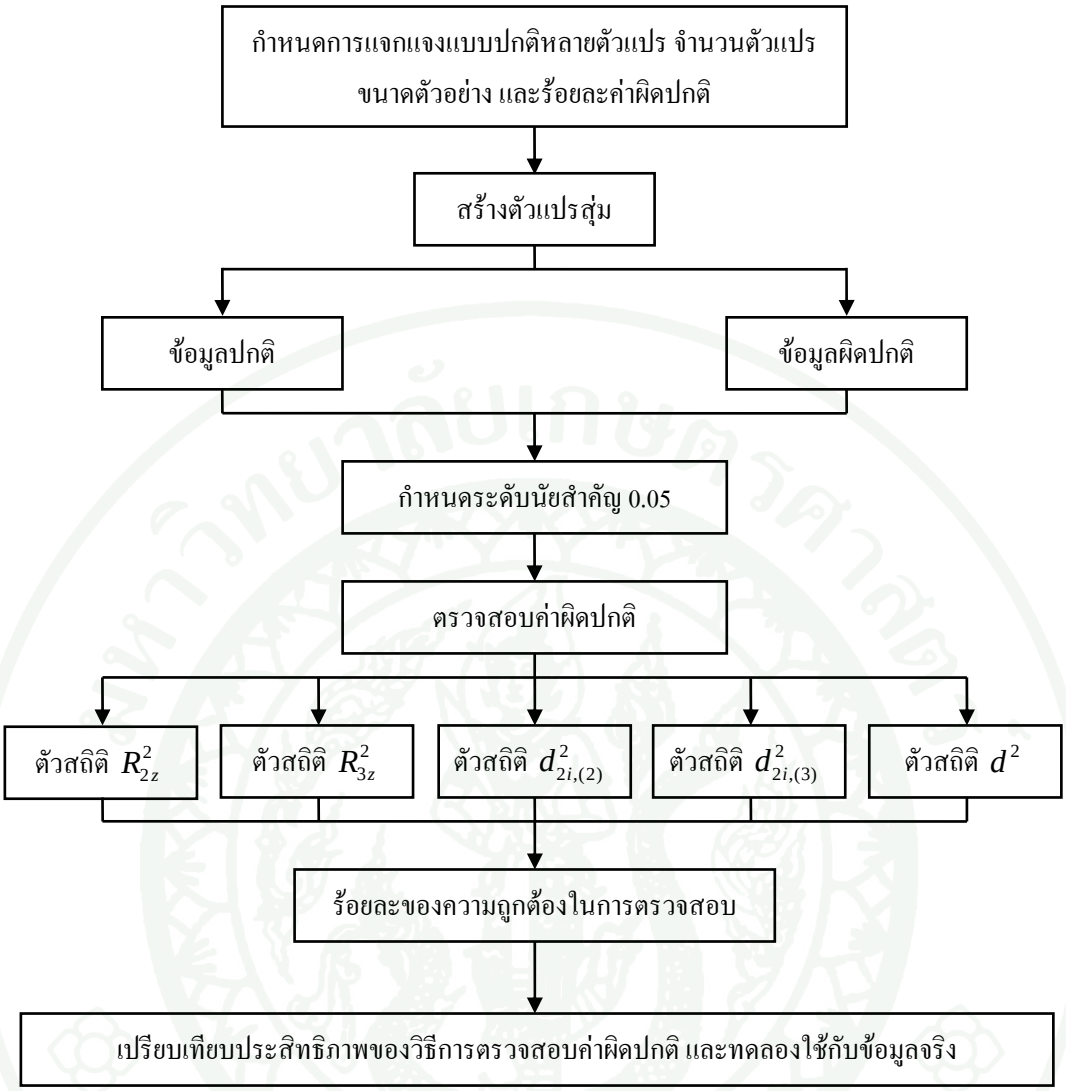
เมื่อนำข้อมูลจากทั้งสองกลุ่มมารวมกันจะได้ชุดข้อมูลที่มีขนาดตัวอย่าง 30 ประกอบด้วยข้อมูลปกติ และข้อมูลผิดปกติปะปนร้อยละ 10 ของข้อมูลทั้งหมด โดยค่าผิดปกติมีลักษณะที่ความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน สำหรับในสถานการณ์อื่นๆ สามารถทำได้ในทำนองเดียวกัน

3. ตรวจสอบค่าผิดปกติโดยใช้ตัวสถิติ R^2_{2z} และ R^2_{3z} จากวิธีที่นำเสนอ ตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จากวิธีที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanobis distance

4. เปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบ ซึ่งชุดข้อมูลจะประกอบด้วยกลุ่มของข้อมูลปกติ และกลุ่มของข้อมูลผิดปกติ ค่าร้อยละของความถูกต้องคำนวณได้จากจำนวนข้อมูลที่ตรวจสอบได้ถูกต้องว่าเป็นข้อมูลปกติ หรือเป็นข้อมูลผิดปกติต่อจำนวนข้อมูลทั้งหมด รวมทั้งทดลองใช้วิธีการตรวจสอบค่าผิดปกติกับข้อมูลจริง เพื่อทำการเปรียบเทียบผลการตรวจสอบที่ได้

5. จำลองข้อมูลโดยทำซ้ำจำนวน 1,000 ครั้ง

จากวิธีการวิจัยข้างต้น สามารถสรุปเป็นวิธีการทดลองเป็นผังงานได้ดังนี้



ภาพที่ 1 แผนผังแสดงขั้นตอนของวิธีการทดลอง

ผลและวิจารณ์

ผล

การวิจัยครั้งนี้ทำการสร้างตัวสถิติ R_{2z}^2 และ R_{3z}^2 ที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรด้วยวิธีการวิเคราะห์องค์ประกอบหลักโดยใช้องค์ประกอบรอง 2 และ 3 องค์ประกอบ เนื่องจากองค์ประกอบรองสามารถนำไปใช้ตรวจสอบค่าผิดปกติที่เกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน จากนั้นทำการเปรียบเทียบวิธีการตรวจสอบค่าผิดปกติของตัวสถิติ R_{2z}^2 และ R_{3z}^2 กับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanbis distance ผลการเปรียบเทียบแบ่งเป็น 2 ส่วน ได้แก่ การเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบค่าผิดปกติกับข้อมูลที่จำลองภายใต้สถานการณ์ที่กำหนดโดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบ และการเปรียบเทียบผลการตรวจสอบค่าผิดปกติกับข้อมูลจริง

ในการสร้างตัวสถิติ R_{2z}^2 และ R_{3z}^2 จากองค์ประกอบรองด้วยวิธีการวิเคราะห์องค์ประกอบหลัก เพื่อใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร ทำได้โดยการนำความรู้ทางทฤษฎีและสมบัติของการแจกแจงแบบปกติมาปรับใช้ ภายใต้ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร โดยนำข้อมูลมาทำการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ ที่มีคู่อันดับของค่าเฉพาะและเวกเตอร์เฉพาะ $(\lambda_1, \boldsymbol{\alpha}_1), (\lambda_2, \boldsymbol{\alpha}_2), ..., (\lambda_p, \boldsymbol{\alpha}_p)$ เมื่อ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ จะได้จำนวนองค์ประกอบเท่ากับจำนวนตัวแปร ซึ่งในแต่ละองค์ประกอบจะมีการแจกแจงแบบปกติ เขียนแทนด้วย $y_i \sim N(0, \lambda_i)$

ตัวสถิติที่ได้จากการพิจารณาองค์ประกอบรอง 2 องค์ประกอบ คือ

$$R_{2z}^2 = \frac{R_2^2}{\lambda_{p-1} + \lambda_p}$$

โดยที่ R_{2z}^2 มีการแจกแจงแบบไคกำลังสอง ที่มีองศาอิสระเท่ากับ 1 เขียนแทนด้วยสัญลักษณ์ $R_{2z}^2 \sim \chi_{(1)}^2$

และตัวสถิติที่ได้จากการพิจาณาองค์ประกอบรอง 3 องค์ประกอบ คือ

$$R_{3z}^2 = \frac{R_3}{\sqrt{\lambda_{p-2} + \lambda_{p-1} + \lambda_p}}$$

โดยที่ R_{3z}^2 มีการแจกแจงแบบไคกำลังสอง ที่มีองศาอิสระเท่ากับ 1 เขียนแทนด้วยสัญลักษณ์ $R_{3z}^2 \sim \chi_{(1)}^2$

ผลจากการเปรียบเทียบประสิทธิภาพในการตรวจสอบค่าผิดปกติของตัวสถิติที่นำเสนอในการศึกษาครั้งนี้ คือ R_{2z}^2 และ R_{3z}^2 กับตัวสถิติอื่นๆ ได้แก่ $d_{2i,(2)}^2$, $d_{2i,(3)}^2$ และ d^2 โดยการจำลองข้อมูลภายใต้สถานการณ์ต่างๆ ที่กำหนด ได้แก่ จำนวนตัวแปร ขนาดข้อมูลตัวอย่าง และร้อยละของค่าผิดปกติ พิจารณาเปรียบเทียบประสิทธิภาพของตัวสถิติต่างๆ จากร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติ สรุปผลได้ดังนี้

ผลการเปรียบเทียบประสิทธิภาพของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p = 5$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 1

ตารางที่ 1 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p=5$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	91.007	87.860	**94.907**	94.320	93.280
	20	83.127	78.247	**86.613**	85.520	83.887
	30	73.380	68.133	**75.513**	74.847	73.133
	40	62.093	58.093	63.467	**64.000**	62.000
40	10	90.900	87.395	**94.980**	94.415	93.045
	20	83.430	78.105	**86.920**	85.590	83.870
	30	73.910	68.120	**76.265**	75.265	73.205
	40	62.815	57.630	**64.400**	64.010	62.090
50	10	91.120	87.584	**94.980**	94.072	92.916
	20	83.196	78.116	**86.968**	85.460	83.860
	30	74.148	68.008	**76.480**	75.340	73.096
	40	62.932	57.332	64.384	**64.392**	62.304
75	10	90.384	86.648	**94.632**	93.576	92.000
	20	83.200	78.376	**87.272**	85.837	83.792
	30	74.067	67.464	**76.435**	74.963	72.675
	40	63.712	57.261	**65.133**	64.680	62.379
100	10	90.562	87.060	**94.840**	93.814	92.484
	20	83.278	78.112	**87.268**	85.706	83.654
	30	75.130	68.214	**77.348**	75.588	73.290
	40	64.112	56.724	**65.502**	64.712	62.444

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 1 พบว่า กรณี $p=5$ เมื่อ $n=30$ ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2 มีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่ถึงร้อยละ 5 สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $n=40$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพมากที่สุดในทุกกรณี และมีค่าประสิทธิภาพใกล้เคียงกับตัวสถิติ $d^2_{2i,(3)}$ ส่วนตัวสถิติ R^2_{2z} มีประสิทธิภาพใกล้เคียงกับตัวสถิติ d^2 โดยตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $n=50$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ซึ่งตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ มีประสิทธิภาพใกล้เคียงกันมาก สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพต่ำกว่าตัวสถิติ d^2 แต่มีร้อยละของความถูกต้องในการตรวจสอบต่ำกว่าไม่เกินร้อยละ 1 ส่วนตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $n=75$ และ $n=100$ ผลการเปรียบเทียบประสิทธิภาพของทั้งสองกรณีจะมีลักษณะเช่นเดียวกัน คือ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี สำหรับตัวสถิติ $R^2_{2z}, d^2_{2i,(2)}, d^2_{2i,(3)}$ และ d^2 จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3 และตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=5$ พบว่า ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุด รองลงมาคือ ตัวสถิติ $d^2_{2i,(3)}$ สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพต่ำกว่าตัวสถิติอื่น โดยมีร้อยละของความถูกต้องในการตรวจสอบต่ำกว่าไม่เกินร้อยละ 5 และตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุด

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p=10$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 2

ตารางที่ 2 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 10$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	88.030	87.897	89.677	89.667	**91.817**
	20	78.407	77.283	79.830	79.163	**80.303**
	30	68.900	67.517	**70.117**	69.270	69.343
	40	59.600	58.000	**60.703**	59.593	58.220
40	10	88.040	88.030	89.710	89.605	**91.373**
	20	78.305	77.178	79.810	79.005	**79.868**
	30	68.825	67.170	**70.445**	69.098	68.383
	40	59.325	57.435	**60.440**	58.848	56.770
50	10	88.150	88.004	89.822	89.822	**90.944**
	20	78.170	76.728	**79.876**	78.926	79.336
	30	68.768	66.944	**70.188**	68.930	67.624
	40	59.292	56.928	**60.302**	58.620	55.938
75	10	87.440	87.217	**89.069**	88.625	89.413
	20	78.060	76.617	**79.951**	78.735	78.872
	30	68.120	65.939	**69.556**	67.819	66.180
	40	58.992	56.645	**60.297**	58.193	54.885
100	10	88.108	87.823	89.722	89.293	**89.777**
	20	77.968	76.566	**79.804**	78.568	78.451
	30	68.741	66.401	**70.209**	68.496	66.665
	40	59.159	56.406	**60.177**	58.003	54.613

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 2 พบว่า กรณี $p = 10$ เมื่อ $n = 30$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อยในทุกกรณี ส่วนตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพแตกต่างจากตัวสถิติอื่น โดยมีร้อยละ

ของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3 ในทุกกรณี และตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติทุกตัวค่อนข้างมาก

เมื่อ $n=40$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อยในทุกกรณี ตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างจากตัวสถิติอื่นไม่เกินร้อยละ 3 แต่ตัวสถิติ R_{3z}^2 มีประสิทธิภาพต่ำที่สุด

เมื่อ $n=50$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น สำหรับร้อยละของค่าผิดปกติระดับอื่นๆ ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2, R_{3z}^2, $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกันโดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3 แต่ตัวสถิติ R_{3z}^2 มีประสิทธิภาพต่ำที่สุด

เมื่อ $n=75$ ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ส่วนตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3 สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น ซึ่งตัวสถิติ R_{3z}^2 ยังคงมีประสิทธิภาพต่ำที่สุด

เมื่อ $n=100$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น สำหรับร้อยละของค่าผิดปกติอื่นๆ ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2, R_{3z}^2, $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ มีประสิทธิภาพใกล้เคียงกันโดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3 โดยตัวสถิติ R_{3z}^2 ยังคงมีประสิทธิภาพต่ำที่สุด

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=10$ พบว่า ในภาพรวมตัวสถิติ R_{3z}^2 มีประสิทธิภาพต่ำที่สุด ส่วนใหญ่ตัวสถิติ $d_{2i,(2)}^2$ มีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n=$ 30 และ 40 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n=50$ และ 100

ส่วนตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ แต่จะมีประสิทธิ์ต่ำกว่าเพียงเล็กน้อย

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p = 12$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 3

ตารางที่ 3 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 12$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	88.620	88.960	90.877	91.767	**95.567**
	20	80.567	80.700	82.533	84.167	**85.133**
	30	71.333	71.553	73.107	**74.033**	72.980
	40	62.233	62.133	63.413	**63.687**	61.447
40	10	90.100	90.320	92.860	92.860	**96.320**
	20	82.110	81.905	84.350	85.780	**87.095**
	30	72.950	72.575	75.305	**75.325**	75.050
	40	63.500	63.275	64.800	**65.350**	63.225
50	10	91.360	90.780	93.920	94.840	**96.120**
	20	82.800	82.380	85.740	87.180	**87.560**
	30	73.660	73.060	75.540	**77.400**	76.440
	40	63.340	62.900	64.880	**66.240**	64.260
75	10	91.613	91.213	94.307	95.307	**95.507**
	20	83.987	83.667	87.000	**88.640**	88.547
	30	73.600	73.320	76.173	**77.333**	76.293
	40	64.360	64.347	66.213	**67.147**	64.840
100	10	92.590	92.590	95.010	**96.000**	95.970
	20	84.500	83.750	87.230	**88.480**	88.290
	30	74.650	73.920	77.620	**78.730**	77.610
	40	64.660	64.460	66.640	**67.710**	65.460

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 3 พบว่า กรณี $p=12$ เมื่อ $n=30$ และ 40 ผลการเปรียบเทียบประสิทธิภาพของทั้งสองกรณีจะมีลักษณะเช่นเดียวกัน คือ ตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าทุกตัวสถิติและใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างจากตัวสถิติอื่นไม่เกินร้อยละ 7 สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ส่วนตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ซึ่งมีประสิทธิภาพสูงกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

เมื่อ $n=50$ ตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าทุกตัวสถิติ โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างจากตัวสถิติอื่นไม่เกินร้อยละ 5 แต่ตัวสถิติ R_{3z}^2 ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ส่วนตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ซึ่งมีประสิทธิภาพสูงกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

เมื่อ $n=75$ ตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าทุกตัวสถิติและใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างจากตัวสถิติอื่นไม่เกินร้อยละ 5 แต่ตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ส่วนตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติระดับอื่นๆ และมีประสิทธิภาพสูงกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

เมื่อ $n=100$ ตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าทุกตัวสถิติและใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างจากตัวสถิติอื่นไม่เกินร้อยละ 5 ซึ่งตัวสถิติ R_{3z}^2 ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี และมีประสิทธิภาพสูงกว่าตัวสถิติ $d_{2i,(2)}^2$ เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=12$ พบว่า ในภาพรวมตัวสถิติ R_{2z}^2 และ R_{3z}^2 มีประสิทธิภาพใกล้เคียงกัน และต่ำกว่าทุกตัวสถิติ โดยจะใกล้เคียงกับตัวสถิติอื่นมากขึ้นเมื่อค่าผิดปกติเพิ่มขึ้น สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ

30 และ 40 ในทุกขนาดตัวอย่าง ส่วนตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n = 30, 40$ และ 50 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n = 75$ โดยตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ $d^2_{2i,(3)}$

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p = 14$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 4

ตารางที่ 4 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 14$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	88.233	88.500	88.533	88.533	**92.033**
	20	79.067	79.333	81.033	80.833	**81.233**
	30	70.067	69.667	69.867	70.367	**70.667**
	40	59.467	59.867	59.200	59.567	**60.067**
40	10	89.700	89.775	90.250	91.525	**93.525**
	20	81.675	80.800	81.125	81.925	**83.325**
	30	70.750	71.100	71.275	71.550	**71.800**
	40	59.975	59.900	60.375	60.550	**60.525**
50	10	90.360	90.720	90.980	92.500	**94.520**
	20	82.220	81.420	83.120	83.680	**84.620**
	30	71.680	71.100	71.940	72.240	**72.820**
	40	59.940	59.700	60.240	60.180	**61.260**
75	10	91.240	90.653	92.013	**93.720**	93.693
	20	82.773	81.427	83.693	**84.613**	84.560
	30	70.893	70.013	71.987	71.613	**72.960**
	40	60.067	59.800	60.760	60.653	**61.560**
100	10	92.400	92.030	93.390	**95.240**	94.130
	20	83.570	82.510	84.480	**85.390**	84.800
	30	72.270	71.320	73.480	73.450	**73.790**
	40	60.870	59.980	61.140	61.020	**61.930**

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 4 พบว่า กรณี $p = 14$ เมื่อ $n = 30$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ส่วนตัวสถิติ $R_{2z}^2, R_{3z}^2, d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ มีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 สำหรับตัวสถิติ R_{2z}^2 และ R_{3z}^2 มีประสิทธิภาพใกล้เคียงกันมาก

เมื่อ $n=40$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ส่วนตัวสถิติ $R_{2z}^2, R_{3z}^2, d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ซึ่งตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุด

เมื่อ $n=50$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ส่วนตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(2)}^2$ จะ มีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ d^2 เพียงเล็กน้อย

เมื่อ $n=75$ และ 100 ผลการเปรียบเทียบประสิทธิภาพของทั้งสองกรณีจะมีลักษณะเช่นเดียวกัน คือ ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(2)}^2$ มีประสิทธิภาพใกล้เคียงกัน โดยตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=14$ พบว่า ในภาพรวมตัวสถิติ R_{2z}^2 และ R_{3z}^2 มีประสิทธิภาพใกล้เคียงกัน และต่ำกว่าตัวสถิติอื่น สำหรับตัวสถิติ d^2 มีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ในทุกขนาดตัวอย่าง และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n=30, 40$ และ 50 ส่วนตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n=75$ และ 100 ในแต่ละกรณีทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p=16$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 5

ตารางที่ 5 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 16$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	87.313	88.393	87.927	89.047	**91.440**
	20	79.467	79.367	79.700	**80.633**	80.467
	30	69.793	69.913	70.273	**70.633**	70.073
	40	60.167	60.113	60.307	**60.427**	60.007
40	10	89.350	89.875	90.550	92.100	**95.100**
	20	85.725	85.825	86.350	**88.600**	87.275
	30	70.950	71.125	71.750	**73.000**	71.900
	40	**61.325**	60.325	61.225	61.100	60.700
50	10	90.020	90.880	90.980	93.180	**95.480**
	20	81.780	82.280	82.960	**85.360**	84.360
	30	71.760	71.820	72.940	**73.720**	73.240
	40	**61.980**	61.260	61.220	60.940	61.360
75	10	90.787	91.627	92.533	94.293	**94.667**
	20	83.067	83.360	84.600	**86.347**	85.747
	30	**74.013**	71.493	73.507	72.400	73.613
	40	**62.147**	60.827	61.800	62.093	61.840
100	10	92.070	92.800	93.320	95.270	**95.590**
	20	83.070	83.830	84.840	**86.820**	85.920
	30	**75.280**	72.660	74.570	73.270	74.470
	40	**62.760**	61.200	62.530	62.710	62.210

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 5 พบว่า กรณี $p = 16$ เมื่อ $n = 30$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติระดับอื่นๆ ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(2)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

เมื่อ $n=40$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d^2_{2i,(3)}$ มีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 โดยทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน และใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

เมื่อ $n=50$ ตัวสถิติ d^2 มีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d^2_{2i,(3)}$ มีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน และใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1

เมื่อ $n=75$ และ 100 ผลการเปรียบเทียบประสิทธิภาพของทั้งสองกรณีจะมีลักษณะเช่นเดียวกัน คือ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน และใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=16$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 5 ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ส่วนตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติระดับอื่นๆ ที่ $n=30$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ที่ $n=40$ และ 50 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n=75$ และ 100 สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ทุกขนาดตัวอย่าง ยกเว้น $n=30$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 ที่ $n=75$ และ 100 ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพแตกต่างกันเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 5

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p = 18$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 6

ตารางที่ 6 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 18$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	87.500	87.500	89.067	89.267	**90.113**
	20	78.267	78.267	78.833	79.667	**80.001**
	30	68.433	68.900	68.600	69.133	**70.133**
	40	59.667	59.500	59.800	60.142	**60.467**
40	10	89.075	88.225	91.300	91.600	**95.050**
	20	78.900	78.525	81.900	82.300	**83.200**
	30	70.100	69.950	70.650	71.300	**71.425**
	40	**60.250**	60.125	59.725	59.950	59.875
50	10	89.540	88.620	92.640	93.080	**95.480**
	20	79.960	80.000	82.480	83.220	**84.580**
	30	**73.000**	72.480	71.420	72.880	70.640
	40	**61.140**	60.500	60.340	60.860	60.020
75	10	90.907	89.613	93.560	94.693	**94.840**
	20	**85.360**	84.653	83.640	84.493	81.173
	30	**73.107**	72.547	71.880	72.560	70.440
	40	**61.547**	60.827	60.707	60.013	61.213
100	10	91.830	90.060	94.880	95.100	**95.300**
	20	**85.790**	83.570	84.740	85.580	82.000
	30	**74.230**	72.380	72.970	71.270	73.780
	40	**61.560**	61.460	60.390	60.100	61.550

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 6 พบว่า กรณี $p=18$ เมื่อ $n=30$ ตัวสถิติ d^2 จะมีประสิทธิภาพมากที่สุดในทุกกรณี ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

เมื่อ $n=40$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพมากที่สุด สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกันมาก

เมื่อ $n=50$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ยังคงมีประสิทธิภาพใกล้เคียงกันมาก

เมื่อ $n=75$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ยังคงมีประสิทธิภาพใกล้เคียงกันมาก

เมื่อ $n=100$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกันมาก

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=18$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องแตกต่างกันไม่เกินร้อยละ 6 ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณีที่ $n=30$ เช่นเดียวกับที่ $n=40$ ยกเว้นเมื่อร้อยละของ

ค่าผิดปกติเท่ากับ 40 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n = 50$ รวมทั้งเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n = 75$ และ 100 สำหรับตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุดเมื่อมีขนาดข้อมูลตัวอย่าง และร้อยละของค่าผิดปกติเพิ่มขึ้น ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $n = 40$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $n = 50$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $n = 75$ และ 100

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $p = 20$ จำแนกตามขนาดข้อมูลตัวอย่าง แสดงดังตารางที่ 7

ตารางที่ 7 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $p = 20$ จำแนกตามขนาดข้อมูลตัวอย่าง

ขนาดข้อมูลตัวอย่าง	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
30	10	86.067	87.067	87.067	88.300	**90.200**
	20	78.000	77.267	78.067	78.533	**80.143**
	30	68.767	69.067	69.367	69.833	**70.002**
	40	59.800	60.200	60.233	**60.367**	60.000
40	10	88.825	88.775	89.625	91.050	**94.225**
	20	80.675	79.700	81.500	**82.350**	82.325
	30	70.575	70.550	71.250	**72.025**	70.875
	40	61.300	60.700	**61.900**	61.875	60.150
50	10	89.820	89.440	91.760	93.080	**96.280**
	20	81.360	81.460	83.340	**85.020**	84.220
	30	**73.840**	71.940	73.140	71.840	73.020
	40	**62.920**	60.860	62.680	61.520	60.980
75	10	91.707	91.280	93.640	**95.093**	95.507
	20	**87.293**	82.813	85.893	83.280	86.920
	30	**80.760**	77.987	80.013	78.173	79.520
	40	**63.547**	61.360	62.053	63.347	62.173
100	10	92.960	92.560	**96.320**	94.790	96.020
	20	**86.900**	83.710	86.420	88.090	83.700
	30	**76.720**	73.470	75.540	73.730	75.050
	40	**64.190**	62.070	63.130	64.050	62.350

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 7 พบว่า กรณี $p = 20$ เมื่อ $n = 30$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 , 20 และ 30 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพไม่แตกต่างกันมาก และมีค่าใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

เมื่อ $n=40$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อค่าผิดปกติเท่ากับ 40 สำหรับตัวสถิติ R^2_{3z} มีประสิทธิภาพต่ำที่สุดในทุกกรณี ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิไม่แตกต่างกันมาก และจะมีค่าใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

เมื่อ $n=50$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพไม่แตกต่างกันมาก และจะมีค่าใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

เมื่อ $n=75$ ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพไม่แตกต่างกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 5

เมื่อ $n=100$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพไม่แตกต่างกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 5

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $p=20$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกัน โดยมีค่าร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 7 และมีประสิทธิภาพใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $n=30$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n=40$ และ 50 ส่วนตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $n=30$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ที่ $n=40$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n=50$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n=75$ สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติ

เท่ากับ 30 และ 40 ที่ $n = 50, 75$ และ 100 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n = 75$ และ 100 ในแต่ละกรณีตัวสถิติทุกตัวมีประสิทธิภาพไม่แตกต่างกันมาก

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $n = 30$ จำแนกตามจำนวนตัวแปร แสดงดังตารางที่ 8

ตารางที่ 8 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $n = 30$ จำแนกตามจำนวนตัวแปร

จำนวนตัวแปร	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R^2_{2z}	R^2_{3z}	$d^2_{2i,(2)}$	$d^2_{2i,(3)}$	d^2
5	10	91.007	87.860	**94.907**	94.320	93.280
	20	83.127	78.247	**86.613**	85.520	83.887
	30	73.380	68.133	**75.513**	74.847	73.133
	40	62.093	58.093	63.467	**64.000**	62.000
10	10	88.030	87.897	89.677	89.667	**91.817**
	20	78.407	77.283	79.830	79.163	**80.303**
	30	68.900	67.517	**70.117**	69.270	69.343
	40	59.600	58.000	**60.703**	59.593	58.220
12	10	88.620	88.960	90.877	91.767	**95.567**
	20	80.567	80.700	82.533	84.167	**85.133**
	30	71.333	71.553	73.107	**74.033**	72.980
	40	62.233	62.133	63.413	**63.687**	61.447
14	10	88.233	88.500	88.533	88.533	**92.033**
	20	79.067	79.333	81.033	80.833	**81.233**
	30	70.067	69.667	69.867	70.367	**70.667**
	40	59.467	59.867	59.200	59.567	**60.067**
16	10	87.313	88.393	87.927	89.047	**91.440**
	20	79.467	79.367	79.700	**80.633**	80.467
	30	69.793	69.913	70.273	**70.633**	70.073
	40	60.167	60.113	60.307	**60.427**	60.007
18	10	87.500	87.500	89.067	89.267	**90.113**
	20	78.267	78.267	78.833	79.667	**80.001**
	30	68.433	68.900	68.600	69.133	**70.133**
	40	59.667	59.500	59.800	60.142	**60.467**
20	10	86.067	87.067	87.067	88.300	**90.200**
	20	78.000	77.267	78.067	78.533	**80.143**
	30	68.767	69.067	69.367	69.833	**70.002**
	40	59.800	60.200	60.233	**60.367**	60.000

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 8 พบว่า กรณี $n = 30$ เมื่อ $p = 5$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุด เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 , 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ

$d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ซึ่งตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมีประสิทธิภาพใกล้เคียงกัน และตัวสถิติ R^2_{2z} และ R^2_{3z} จะมีประสิทธิภาพใกล้เคียงกัน แต่ตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p = 10$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ซึ่งตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 และตัวสถิติ R^2_{2z} และ R^2_{3z} ยังคงมีประสิทธิภาพใกล้เคียงกันมากเช่นเดียวกัน แต่ตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p = 12$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุด โดยตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 และตัวสถิติ R^2_{2z} และ R^2_{3z} ยังคงมีประสิทธิภาพใกล้เคียงกันมากเช่นเดียวกัน

เมื่อ $p = 14$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

เมื่อ $p = 16$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด โดยตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมีประสิทธิภาพใกล้เคียงกันมาก และตัวสถิติ R^2_{2z} และ R^2_{3z} ก็จะมีประสิทธิภาพใกล้เคียงกันมากเช่นเดียวกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1

เมื่อ $p = 18$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี โดยตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมีประสิทธิภาพใกล้เคียงกันมาก และตัวสถิติ R^2_{2z} และ R^2_{3z} จะมีประสิทธิภาพใกล้เคียงกันมากเช่นเดียวกัน และเท่ากันในบางกรณี

เมื่อ $p = 20$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ซึ่งในแต่ละกรณีทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกันมาก

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $n = 30$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 7 ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณีเมื่อ $p = 14$ และ 18 และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $p = 10$ และ 12 เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $p = 20$ ส่วนตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อ ร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $p = 5$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 10$ สำหรับตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $p = 5$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 12$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p = 16$ โดยตัวสถิติ R^2_{2z} และ R^2_{3z} จะมีประสิทธิภาพไม่แตกต่างกันมาก และต่ำกว่าตัวสถิติอื่น

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $n = 40$ จำแนกตามจำนวนตัวแปร แสดงดังตารางที่ 9

ตารางที่ 9 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $n = 40$ จำแนกตามจำนวนตัวแปร

จำนวนตัวแปร	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
5	10	90.900	87.395	**94.980**	94.415	93.045
	20	83.430	78.105	**86.920**	85.590	83.870
	30	73.910	68.120	**76.265**	75.265	73.205
	40	62.815	57.630	**64.400**	64.010	62.090
10	10	88.040	88.030	89.710	89.605	**91.373**
	20	78.305	77.178	79.810	79.005	**79.868**
	30	68.825	67.170	**70.445**	69.098	68.383
	40	59.325	57.435	**60.440**	58.848	56.770
12	10	90.100	90.320	92.850	92.860	**96.320**
	20	82.110	81.905	84.350	85.780	**87.095**
	30	72.950	72.575	75.305	**75.325**	75.050
	40	63.500	63.275	64.800	**65.350**	63.225
14	10	89.700	89.775	90.250	91.525	**93.525**
	20	81.675	80.800	81.125	81.925	**83.325**
	30	70.750	71.100	71.275	71.550	**71.800**
	40	59.975	59.900	60.375	60.550	**60.525**
16	10	89.350	89.875	90.550	92.100	**95.100**
	20	85.725	85.825	86.350	**88.600**	87.275
	30	70.950	71.125	71.750	**73.000**	71.900
	40	**61.325**	60.325	61.225	61.100	60.700
18	10	89.075	88.225	91.300	91.600	**95.050**
	20	78.900	78.525	81.900	82.300	**83.200**
	30	70.100	69.950	70.650	71.300	**71.425**
	40	**60.250**	60.125	59.725	59.950	59.875
20	10	88.825	88.775	89.625	91.050	**94.225**
	20	80.675	79.700	81.500	**82.350**	82.325
	30	70.575	70.550	71.250	**72.025**	70.875
	40	61.300	60.700	**61.900**	61.875	60.150

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 9 พบว่า กรณี $n = 40$ เมื่อ $p = 5$ ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี และมีประสิทธิภาพใกล้เคียงกับตัวสถิติ $d_{2i,(3)}^2$ โดยมีร้อยละของความถูกต้องใน

การตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ d^2 สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นในทุกกรณี

เมื่อ $p=10$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

เมื่อ $p=12$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 ส่วนตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(3)}^2$ เพียงเล็กน้อย

เมื่อ $p=14$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี โดยจะมีประสิทธิภาพแตกต่างกับตัวสถิติอื่นน้อยลง เมื่อร้อยละของค่าความผิดปกติเพิ่มขึ้น

เมื่อ $p=16$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ซึ่งตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย

เมื่อ $p=18$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R_{2z}^2 เพียงเล็กน้อย โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 ส่วนตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(3)}^2$ เพียงเล็กน้อยเช่นเดียวกัน

เมื่อ $p=20$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ส่วนตัวสถิติ R_{2z}^2

และ R_{3z}^2 จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 แต่ทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกันเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $n = 40$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องแตกต่างกันไม่เกินร้อยละ 7 ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณีเมื่อ $p =$ 14 และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $p = 10$ และ 12 เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 16$ และ 20 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $p = 18$ โดยจะมีประสิทธิภาพลดลงกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น ส่วนตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณีที่ $p =$ 5 และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p =$ 10 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 เมื่อ $p =$ 20 สำหรับตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p =$ 12 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ที่ $p =$ 16 และ 20 โดยตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $p =$ 16 และ 20 ซึ่งในแต่ละกรณีจะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ R_{3z}^2

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $n = 50$ จำแนกตามจำนวนตัวแปร แสดงดังตารางที่ 10

ตารางที่ 10 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $n = 50$ จำแนกตามจำนวนตัวแปร

จำนวนตัวแปร	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R^2_{2z}	R^2_{3z}	$d^2_{2i,(2)}$	$d^2_{2i,(3)}$	d^2
5	10	91.120	87.584	**94.980**	94.072	92.916
	20	83.196	78.116	**86.968**	85.460	83.860
	30	74.148	68.008	**76.480**	75.340	73.096
	40	62.932	57.332	64.384	**64.392**	62.304
10	10	88.150	88.004	89.822	89.822	**90.944**
	20	78.170	76.728	**79.876**	78.926	79.336
	30	68.768	66.944	**70.188**	68.930	67.624
	40	59.292	56.928	**60.302**	58.620	55.938
12	10	91.360	90.780	93.920	94.840	**96.120**
	20	82.800	82.380	85.740	87.180	**87.560**
	30	73.660	73.060	75.540	**77.400**	76.440
	40	63.340	62.900	64.880	**66.240**	64.260
14	10	90.360	90.720	90.980	92.500	**94.520**
	20	82.220	81.420	83.120	83.680	**84.620**
	30	71.680	71.100	71.940	72.240	**72.820**
	40	59.940	59.700	60.240	60.180	**61.260**
16	10	90.020	90.880	90.980	93.180	**95.480**
	20	81.780	82.280	82.960	**85.360**	84.360
	30	71.760	71.820	72.940	**73.720**	73.240
	40	**61.980**	61.260	61.220	60.940	61.360
18	10	89.540	88.620	92.640	93.080	**95.480**
	20	79.960	80.000	82.480	83.220	**84.580**
	30	**73.000**	72.480	71.420	72.880	70.640
	40	**61.140**	60.500	60.340	60.860	60.020
20	10	89.820	89.440	91.760	93.080	**96.280**
	20	81.360	81.460	83.340	**85.020**	84.220
	30	**73.840**	71.940	73.140	71.840	73.020
	40	**62.920**	60.860	62.680	61.520	60.980

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 10 พบว่า กรณี $n = 50$ เมื่อ $p = 5$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ยกเว้นเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด

ส่วนตัวสถิติ d^2 จะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ R_{2z}^2 โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี โดยมีร้อยละของความถูกต้องในการตรวจสอบต่ำกว่าตัวสถิติอื่นถึงร้อยละ 7

เมื่อ $p=10$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2 และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน สำหรับตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ตัวสถิติ d^2 จะมีประสิทธิภาพต่ำที่สุด

เมื่อ $p=12$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 1 สำหรับตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d_{2i,(3)}^2$

เมื่อ $p=14$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน เช่นเดียวกับตัวสถิติ R_{2z}^2 และ R_{3z}^2 ที่มีประสิทธิภาพใกล้เคียงกัน

เมื่อ $p=16$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ซึ่งตัวสถิติ R_{2z}^2, R_{3z}^2 และ $d_{2i,(2)}^2$ จะมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2

เมื่อ $p=18$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุด โดยตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะมีประสิทธิภาพใกล้เคียงกัน เช่นเดียวกับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ที่มีประสิทธิภาพใกล้เคียงกัน

เมื่อ $p=20$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ซึ่งตัวสถิติ R^2_{2z} และ R^2_{3z} ยังคงมีประสิทธิภาพใกล้เคียงกัน เช่นเดียวกับตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ ที่มีประสิทธิภาพใกล้เคียงกัน

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $n=50$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องแตกต่างกันไม่เกินร้อยละ 7 ซึ่งตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดในทุกกรณีที่ $p=$ 14 และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p=10$, 16 และ 20 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $p=12$ และ 18 โดยจะมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น ส่วนตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อ ร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $p=5$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p=10$ สำหรับตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $p=5$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p=12$ เมื่อ ร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ที่ $p=16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $p=20$ ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $p=16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p=18$ และ 20 และจะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ R^2_{3z}

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $n=75$ จำแนกตามจำนวนตัวแปร แสดงดังตารางที่ 11

ตารางที่ 11 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $n = 75$ จำแนกตามจำนวนตัวแปร

จำนวนตัวแปร	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R^2_{2z}	R^2_{3z}	$d^2_{2i,(2)}$	$d^2_{2i,(3)}$	d^2
5	10	90.384	86.648	**94.632**	93.576	92.000
	20	83.200	78.376	**87.272**	85.837	83.792
	30	74.067	67.464	**76.435**	74.963	72.675
	40	63.712	57.261	**65.133**	64.680	62.379
10	10	87.440	87.217	**89.069**	88.625	89.413
	20	78.060	76.617	**79.951**	78.735	78.872
	30	68.120	65.939	**69.556**	67.819	66.180
	40	58.992	56.645	**60.297**	58.193	54.885
12	10	91.613	91.213	94.307	95.307	**95.507**
	20	83.987	83.667	87.000	**88.640**	88.547
	30	73.600	73.320	76.173	**77.333**	76.293
	40	64.360	64.347	66.213	**67.147**	64.840
14	10	91.240	90.653	92.013	**93.720**	93.693
	20	82.773	81.427	83.693	**84.613**	84.560
	30	70.893	70.013	71.987	71.613	**72.960**
	40	60.067	59.800	60.760	60.653	**61.560**
16	10	90.787	91.627	92.533	94.293	**94.667**
	20	83.067	83.360	84.600	**86.347**	85.747
	30	**74.013**	71.493	73.507	72.400	73.613
	40	**62.147**	60.827	61.800	62.093	61.840
18	10	90.907	89.613	93.560	94.693	**94.840**
	20	**85.360**	84.653	83.640	84.493	81.173
	30	**73.107**	72.547	71.880	72.560	70.440
	40	**61.547**	60.827	60.707	60.013	61.213
20	10	91.707	91.280	93.640	**95.093**	95.507
	20	**87.293**	82.813	85.893	83.280	86.920
	30	**80.760**	77.987	80.013	78.173	79.520
	40	**63.547**	61.360	62.053	63.347	62.173

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 11 พบว่า กรณี $n = 75$ เมื่อ $p = 5$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี โดยตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุด

เมื่อ $p=10$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี โดยตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงกว่าตัวสถิติ R^2_{3z} เล็กน้อย ส่วนตัวสถิติ d^2 จะมีประสิทธิภาพใกล้เคียงกับตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ แต่จะมีประสิทธิภาพลดลงกว่าตัวสถิติอื่นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

เมื่อ $p=12$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุด โดยตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงกว่าตัวสถิติ R^2_{3z} เล็กน้อย และตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพต่ำกว่าตัวสถิติ $d^2_{2i,(3)}$ เพียงเล็กน้อยเช่นเดียวกัน

เมื่อ $p=14$ ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุด โดยตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงกว่าตัวสถิติ R^2_{3z} เล็กน้อย สำหรับตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ ยังคงมีประสิทธิภาพใกล้เคียงกัน

เมื่อ $p=16$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด สำหรับตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ส่วนตัวสถิติ R^2_{3z} และ $d^2_{2i,(2)}$ จะมีประสิทธิภาพใกล้เคียงกัน

เมื่อ $p=18$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติมีค่ามากขึ้น

เมื่อ $p=20$ ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $n=75$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องแตกต่างกันไม่เกินร้อยละ 8 และจะมีประสิทธิภาพใกล้เคียงกันมากขึ้นเมื่อขนาดข้อมูลตัวอย่างเพิ่มขึ้น ซึ่งตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ที่ $p=5$ และ 10 ส่วนตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด

ในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p = 12$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $p = 14$ เมื่อ ร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $p = 16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 20$ สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 12, 16$ และ 18 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 14$ ซึ่งตัวสถิติ R_{2z}^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p = 18$ และ 20 ในแต่ละกรณีทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกันมากขึ้น เมื่อร้อยละของค่าผิดปกติมีค่าเพิ่มขึ้น โดยตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำกว่าตัวสถิติอื่นในหลายกรณี

ผลการเปรียบเทียบประสิทธิภาพของของตัวสถิติในการตรวจสอบค่าผิดปกติ โดยพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบของตัวสถิติเมื่อกำหนดร้อยละของค่าผิดปกติต่างๆ เมื่อ $n = 100$ จำแนกตามจำนวนตัวแปร แสดงดังตารางที่ 12

ตารางที่ 12 ร้อยละของความถูกต้องในการตรวจสอบค่าผิดปกติของตัวสถิติต่างๆ เมื่อ $n = 100$ จำแนกตามจำนวนตัวแปร

จำนวนตัวแปร	ร้อยละของค่าผิดปกติ	ร้อยละของความถูกต้องในการตรวจสอบ				
		R^2_{2z}	R^2_{3z}	$d^2_{2i,(2)}$	$d^2_{2i,(3)}$	d^2
5	10	90.562	87.060	**94.840**	93.814	92.484
	20	83.278	78.112	**87.268**	85.706	83.654
	30	75.130	68.214	**77.348**	75.588	73.290
	40	64.112	56.724	**65.502**	64.712	62.444
10	10	88.108	87.823	89.722	89.293	**89.777**
	20	77.968	76.566	**79.804**	78.568	78.451
	30	68.741	66.401	**70.209**	68.496	66.665
	40	59.159	56.406	**60.177**	58.003	54.613
12	10	92.590	92.590	95.010	**96.000**	95.970
	20	84.500	83.750	87.230	**88.480**	88.290
	30	74.650	73.920	77.620	**78.730**	77.610
	40	64.660	64.460	66.640	**67.710**	65.460
14	10	92.400	92.030	93.390	**95.240**	94.130
	20	83.570	82.510	84.480	**85.390**	84.800
	30	72.270	71.320	73.480	73.450	**73.790**
	40	60.870	59.980	61.140	61.020	**61.930**
16	10	92.070	92.800	93.320	95.270	**95.590**
	20	83.070	83.830	84.840	**86.820**	85.920
	30	**75.280**	72.660	74.570	73.270	74.470
	40	**62.760**	61.200	62.530	62.710	62.210
18	10	91.830	90.060	94.880	95.100	**95.300**
	20	**85.790**	83.570	84.740	85.580	82.000
	30	**74.230**	72.380	72.970	71.270	73.780
	40	**61.560**	61.460	60.390	60.100	61.550
20	10	92.960	92.560	**96.320**	94.790	96.020
	20	**86.900**	83.710	86.420	88.090	83.700
	30	**76.720**	73.470	75.540	73.730	75.050
	40	**64.190**	62.070	63.130	64.050	62.350

หมายเหตุ ตัวหนา หมายถึง ร้อยละของความถูกต้องในการตรวจสอบมากที่สุด

จากตารางที่ 12 พบว่า กรณี $n=100$ เมื่อ $p=5$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี และมีประสิทธิภาพสูงกว่าตัวสถิติ $d^2_{2i,(3)}$ โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p=10$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุด และมีประสิทธิภาพใกล้เคียงกับตัวสถิติ $d^2_{2i,(3)}$ ส่วนตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p=12$ ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี และมีประสิทธิภาพสูงกว่าตัวสถิติ $d^2_{2i,(2)}$ โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 2 ส่วนตัวสถิติ R^2_{2z} และ R^2_{3z} จะมีประสิทธิภาพใกล้เคียงกัน แต่ตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p=14$ ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำกว่าตัวสถิติ R^2_{2z} เพียงเล็กน้อย โดยยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p=16$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ส่วนตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี

เมื่อ $p=18$ ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด โดยทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

เมื่อ $p=20$ ตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 แต่เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุด ส่วนตัวสถิติ R^2_{3z} ยังคงมีประสิทธิภาพต่ำที่สุดในทุกกรณี สำหรับตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จะมี

ประสิทธิภาพใกล้เคียงกัน โดยมีร้อยละของความถูกต้องในการตรวจสอบแตกต่างกันไม่เกินร้อยละ 3

สรุปผลการเปรียบเทียบประสิทธิภาพกรณี $n = 100$ พบว่า ในภาพรวมทุกตัวสถิติมีประสิทธิภาพใกล้เคียงกันมาก โดยมีร้อยละของความถูกต้องแตกต่างกันไม่เกินร้อยละ 5 ซึ่งตัวสถิติ $d^2_{2i,(2)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ที่ $p = 5$ และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p = 10$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 20$ ส่วนตัวสถิติ $d^2_{2i,(3)}$ จะมีประสิทธิภาพสูงที่สุดในทุกกรณี ที่ $p = 12$ และในบางกรณี ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $p = 14$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $p = 16$ สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $p = 10, 16$ และ 18 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 14$ นอกนั้นจะเป็นกรณีที่ตัวสถิติ R^2_{2z} มีประสิทธิภาพสูงที่สุด ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $p = 16$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $p = 18$ และ 20 โดยตัวสถิติ R^2_{3z} จะมีประสิทธิภาพต่ำที่สุดในหลายกรณี โดยทุกตัวสถิติจะมีประสิทธิภาพใกล้เคียงกันมากขึ้นเมื่อร้อยละของค่าผิดปกติเพิ่มขึ้น

สำหรับการนำวิธีการตรวจสอบค่าผิดปกติมาใช้กับข้อมูลจริง ซึ่งเป็นข้อมูลทุติยภูมิจากบริษัทเงินทุนแห่งหนึ่งที่นำมาใช้ในการพัฒนาตัวแบบประเมินความเสี่ยงด้านสินเชื่อ (credit rating model) เพื่อใช้ประเมินโอกาสที่ลูกหนี้จะผิดชำระหนี้ (probability of default) ในการจัดอันดับความน่าเชื่อถือของผู้ขอกู้ (ลูกหนี้) ซึ่งเป็นองค์ประกอบหนึ่งสำหรับพิจารณาอนุมัติ หรือไม่อนุมัติเงินกู้ให้กับผู้ขอเงินกู้แต่ละราย ข้อมูลชุดนี้ประกอบด้วยตัวแปรอัตราส่วนทางการเงิน (financial ratio) จำนวน 10 ตัวแปร โดยมีจำนวนข้อมูลตัวอย่าง 65 ชุดข้อมูล ซึ่งแสดงดังตารางผนวกที่ 1 โดยนำตัวสถิติ R^2_{2z} และ R^2_{3z} จากวิธีที่นำเสนอ ตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จากวิธีที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanobis distance มาทำการตรวจสอบเพื่อหาค่าผิดปกติจากข้อมูลดังกล่าว ซึ่งได้ผลการตรวจสอบแสดงดังตารางที่ 13

ตารางที่ 13 แสดงผลการตรวจสอบค่าผิดปกติในแต่ละชุดข้อมูล จำแนกตามตัวสถิติ

ชุดข้อมูลที่	R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
1					
2	*	*	*		*
3			*	*	
4					
5					
6					
7					
8	*	*	*	*	*
9				*	
10					
11					
12			*		
13					
14					*
15					
16					
17					
18					

ตารางที่ 13 (ต่อ)

ชุดข้อมูลที่	R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
19					
20					
21					
22					*
23					
24					
25					
26					
27					
28					*
29					
30					
31					
32					*
33					
34					
35					
36					

ตารางที่ 13 (ต่อ)

ชุดข้อมูลที่	R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
37					
38	*	*	*	*	
39					
40					
41					
42					
43					
44					
45					*
46					
47					
48					
49					
50					
51					
52					
53	*	*	*	*	*
54					

ตารางที่ 13 (ต่อ)

ชุดข้อมูลที่	R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
55					
56					
57					*
58					*
59					
60					
61					*
62					
63					
64					
65					

หมายเหตุ * หมายถึง ชุดข้อมูลที่ถูกตรวจสอบพบว่าเป็นค่าผิดปกติ

จากตารางที่ 13 พบว่า ผลการตรวจสอบค่าผิดปกติของข้อมูลดังกล่าวด้วยตัวสถิติ R_{2z}^2 และ R_{3z}^2 ตรวจพบค่าผิดปกติ 4 ค่า ได้แก่ ชุดข้อมูลที่ 2, 8, 38 และ 53 ส่วนตัวสถิติ $d_{2i,(2)}^2$ ตรวจพบค่าผิดปกติ 6 ค่า ได้แก่ ชุดข้อมูลที่ 2, 3, 8, 12, 35 และ 53 สำหรับตัวสถิติ $d_{2i,(3)}^2$ ตรวจพบค่าผิดปกติ 5 ค่า ได้แก่ ชุดข้อมูลที่ 3, 8, 9, 38 และ 53 ตัวสถิติ d^2 ตรวจพบค่าผิดปกติมากถึง 11 ค่า ได้แก่ ชุดข้อมูลที่ 2, 8, 14, 22, 28, 32, 45, 53, 57, 58 และ 61 โดยเมื่อนำผลการตรวจสอบดังกล่าวมาสรุปรวมแสดงดังตารางที่ 14

ตารางที่ 14 แสดงการเปรียบเทียบชุดข้อมูลที่ตรวจพบเป็นค่าผิดปกติ จำแนกตามตัวสถิติ

ชุดข้อมูลที่				
R_{2z}^2	R_{3z}^2	$d_{2i,(2)}^2$	$d_{2i,(3)}^2$	d^2
2	2	2	3	2
8	8	3	8	8
38	38	8	9	14
53	53	12	38	22
		35	53	28
		53		32
				45
				53
				57
				58
				61

จากตารางที่ 14 พบว่า ตัวสถิติ R_{2z}^2 สามารถตรวจพบค่าผิดปกติจำนวน 4 ชุดข้อมูล ซึ่งได้ผลเช่นเดียวกับตัวสถิติ R_{3z}^2 และผลการตรวจสอบใกล้เคียงเช่นเดียวกับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ ตรวจพบค่าผิดปกติจำนวน 6 และ 5 ชุดข้อมูล โดยส่วนใหญ่เป็นชุดข้อมูลเดียวกัน สำหรับตัวสถิติ d^2 ตรวจสอบค่าผิดปกติถึงจำนวน 11 ชุดข้อมูล และในบางชุดข้อมูลไม่สามารถตรวจพบได้จากตัวสถิติอื่น ซึ่งน่าจะเป็นผลมาจากการที่ตัวสถิติ d^2 ของวิธี Mahalanobis distance จะคำนวณจากการที่ใช้ชุดข้อมูลทุกตัว และมีค่าเท่ากับผลรวมกำลังสองของค่ามาตรฐานในทุกองค์ประกอบอีกด้วย (Jobson, 1992) จึงเป็นตัวสถิติที่ได้มาจากการใช้ทุกองค์ประกอบนั่นเอง ส่งผลทำให้ตัวสถิติ d^2 สามารถเก็บรายละเอียดของข้อมูลได้ดีกว่าตัวสถิติ R_{2z}^2, R_{3z}^2, $d_{2i,(2)}^2$ และ

$d^2_{2i,(3)}$ ซึ่งเป็นตัวสถิติที่ใช้เพียงบางองค์ประกอบเท่านั้น สำหรับการใช้วิธี Mahalanobis distance ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร อาจจะพบปัญหา swamping effect และ masking effect (Penny and Jolliffe, 2001) และยังมีการคำนวณที่ค่อนข้างยุ่งยาก ซับซ้อน

อย่างไรก็ตามตัวสถิติ R^2_{2z} และ R^2_{3z} เป็นตัวสถิติที่คำนวณได้ง่าย รวมทั้งยังมีประสิทธิภาพไม่แตกต่างจากตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ มากนัก และเป็นตัวสถิติที่ใช้เฉพาะองค์ประกอบรองเมื่อมีจำนวนเท่ากันอีกด้วย

วิจารณ์

จากตัวสถิติ R^2_{2z} และ R^2_{3z} ที่ใช้ในการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยใช้องค์ประกอบรอง 2 และ 3 องค์ประกอบ ตามลำดับ กรณีที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน ซึ่งรูปแบบการแจกแจงของตัวสถิติทั้งสองจะมีการแจกแจงแบบไคกำลังสอง ที่มีองศาอิสระเท่ากับ 1 มีรูปแบบการคำนวณที่ไม่ยุ่งยาก สามารถคำนวณได้ง่ายและรวดเร็ว โดยไม่ต้องนำสารสนเทศหรือรายละเอียดของข้อมูลทั้งหมดมาวิเคราะห์นอกจากองค์ประกอบรอง 2-3 องค์ประกอบเท่านั้น ซึ่งจะแตกต่างจากตัวสถิติ d^2 จากวิธี Mahalanobis distance ที่ไม่มีการตัดรายละเอียดของข้อมูลทิ้ง และมีการคำนวณที่ยุ่งยากซับซ้อน สำหรับตัวสถิติ $d^2_{2i,(2)}$ และ $d^2_{2i,(3)}$ จากวิธีที่นำเสนอโดย Hawkins จะใช้องค์ประกอบรอง 2-3 องค์ประกอบเช่นเดียวกัน แต่มีรูปแบบของตัวสถิติและการแจกแจงแตกต่างจากตัวสถิติที่ได้นำเสนอ

จากการเปรียบเทียบประสิทธิภาพของตัวสถิติทั้งหมด พบว่า ตัวสถิติแต่ละตัวจะเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติในแต่ละสถานการณ์ที่แตกต่างกัน ดังนี้

1. ตัวสถิติ R^2_{2z} จะมีประสิทธิภาพสูงที่สุดในบางกรณี ได้แก่ กรณี $p = 16$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $n = 40, 50, 75$ และ 100 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 30 ที่ $n = 75$ และ 100 สำหรับกรณี $p = 18$ ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $n = 40, 50, 75$ และ 100 เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 ที่ $n = 50, 75$ และ 100 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n = 75$ และ 100 ส่วนกรณี $p = 20$ ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 40 ที่ $n = 50, 75$ และ 100 เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 ที่ $n = 50, 75$ และ 100 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n = 75$ และ 100 ดังนั้น ตัวสถิติ R^2_{2z} จึงเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อร้อยละของค่าผิดปกติตั้งแต่ 30 ขึ้นไป ที่ $p \geq 16$ และ $n \geq 50$

2. ตัวสถิติ R_{3z}^2 จะมีประสิทธิภาพต่ำที่สุดในหลายกรณี แต่มีประสิทธิภาพไม่แตกต่างกับตัวสถิติ R_{2z}^2 มากนัก ดังนั้น ตัวสถิติ R_{3z}^2 จึงยังไม่เหมาะที่จะนำไปใช้ในการตรวจสอบค่าผิดปกติเท่าใดนัก

3. ตัวสถิติ $d_{2i,(2)}^2$ จะมีประสิทธิภาพสูงเป็นส่วนใหญ่ในกรณี $p=5$ สำหรับกรณี $p=10$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติ 30 และ 40 ที่ทุกขนาดข้อมูลตัวอย่าง และในบางกรณีเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ดังนั้น ตัวสถิติ $d_{2i,(2)}^2$ จึงเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อร้อยละของค่าผิดปกติตั้งแต่ 30 ขึ้นไป ที่ $p \le 10$

4. ตัวสถิติ $d_{2i,(3)}^2$ จะมีประสิทธิภาพสูงที่สุดเป็นส่วนใหญ่ ได้แก่ กรณี $p=12$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 30 และ 40 ที่ $n=40$ และ 50 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20, 30 และ 40 ที่ $n=75$ และ 100 ส่วนกรณี $p=14$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n=75$ และ 100 และในกรณี $p=16$ ได้แก่ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 , 30 และ 40 ที่ $n=30$ เมื่อร้อยละของค่าผิดปกติเท่ากับ 20 และ 30 ที่ $n=40$ และ 50 และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $n=75$ และ 100 ดังนั้น ตัวสถิติ $d_{2i,(3)}^2$ จึงเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อร้อยละของค่าผิดปกติตั้งแต่ 30 ขึ้นไป ที่ $12 \le p \le 14$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 20 ที่ $p=16$ ในทุกขนาดข้อมูลตัวอย่าง

5. ตัวสถิติ d^2 จะมีประสิทธิภาพสูงที่สุดเป็นส่วนใหญ่ ได้แก่ กรณี $p=14$ เมื่อร้อยละของค่าผิดปกติในทุกกรณี ยกเว้นเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 และ 20 ที่ $n=75$ และ 100 สำหรับในบางกรณี ได้แก่ กรณี $p=16$ และ 18 เมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ทุกขนาดข้อมูลตัวอย่าง และในกรณี $p=20$ จะมีประสิทธิภาพสูงที่สุดเมื่อร้อยละของค่าผิดปกติเท่ากับ 10, 20 และ 30 ที่ $n=30$ และเมื่อร้อยละของค่าผิดปกติเท่ากับ 10 ที่ $n=40$ และ 50 ดังนั้น ตัวสถิติ d^2 จึงเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อ $p=14$ และ $p>14$ แต่ร้อยละของค่าผิดปกติไม่เกิน 30

6. สำหรับตัวสถิติ d^2 จะมีประสิทธิภาพสูงกว่าตัวสถิติอื่นในหลายกรณี ซึ่งเป็นผลมาจากพื้นฐานของตัวสถิติจะคำนวณจากระยะห่างระหว่างข้อมูลทุกตัวกับค่าเฉลี่ย โดยไม่มีการละทิ้งรายละเอียดของข้อมูล แต่การคำนวณก็จะมีความยุ่งยากมาก

สำหรับการนำวิธีการตรวจสอบค่าผิดปกติมาใช้กับข้อมูลของบริษัทเงินทุนแห่งหนึ่งในด้านสินเชื่อ ซึ่งผลที่ได้ในแต่ละตัวสถิติจะมีความแตกต่างกันบ้าง แต่ตัวสถิติ R_{2z}^2 จะให้ผลการตรวจสอบเช่นเดียวกับตัวสถิติ R_{3z}^2 ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะให้ผลการตรวจสอบทำนองเดียวกัน แต่อาจมีเพียงบางชุดข้อมูลที่แตกต่างกันเพียงเล็กน้อย สำหรับตัวสถิติ d^2 จะให้ผลการตรวจสอบพบค่าผิดปกติเป็นจำนวนมาก ซึ่งอาจจะเป็นผลจากที่ตัวสถิติคำนวณมาจากทุกองค์ประกอบ อย่างไรก็ตามการพิจารณาว่าควรจะสรุปผลว่าข้อมูลชุดใดควรจะเป็นค่าผิดปกตินั้นในทางปฏิบัติทำได้ยาก เนื่องจากในสภาพการณ์จริงนักสถิติจะไม่สามารถทราบได้ล่วงหน้าว่าข้อมูลชุดใดเป็นค่าที่ผิดปกติอย่างแท้จริง ดังนั้นวิธีการตรวจสอบเหล่านี้สามารถใช้เป็นแนวทางในการตัดสินใจเพื่อสรุปว่าข้อมูลชุดใดควรจะเป็นค่าผิดปกติ ซึ่งโดยทั่วไปแล้วชุดข้อมูลที่พบว่าเป็นค่าผิดปกติในวิธีการตรวจสอบที่แตกต่างกัน มักจะมีแนวโน้มว่าข้อมูลชุดนั้นน่าจะมีแนวโน้มเป็นค่าผิดปกตินั่นเอง

สรุปและข้อเสนอแนะ

สรุป

การศึกษาวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร โดยใช้องค์ประกอบรอง กรณีที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน ตัวสถิติดังกล่าว ได้แก่ ตัวสถิติ R_{2z}^2 และ R_{3z}^2 ซึ่งได้มาจากการนำหลักการ และทฤษฎีที่เกี่ยวข้องกับวิธีการวิเคราะห์องค์ประกอบหลัก รวมทั้งสมบัติของการแจกแจงแบบปกติ โดยตัวสถิติ R_{2z}^2 จะใช้องค์ประกอบรอง 2 องค์ประกอบ และตัวสถิติ R_{3z}^2 จะใช้องค์ประกอบรอง 3 องค์ประกอบ ตัวสถิติทั้งสองจะมีการแจกแจงแบบไคกำลังสอง ที่มีองศาอิสระเท่ากับ 1 เมื่อได้ตัวสถิติดังกล่าวแล้วจะทำการเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปรของตัวสถิติ R_{2z}^2 และ R_{3z}^2 จากวิธีที่ได้นำเสนอ กับตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จากวิธีที่นำเสนอโดย Hawkins และตัวสถิติ d^2 จากวิธี Mahalanobis distance โดยการจำลองข้อมูลภายใต้สถานการณ์ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร กำหนดจำนวนตัวแปร $p = 5$, 10(2)20 ขนาดข้อมูลตัวอย่าง $n =$ 30, 40, 50, 75, 100 และร้อยละของค่าผิดปกติ 4 ระดับ ได้แก่ 10, 20, 30 และ 40 รวมทั้งหมด 140 สถานการณ์ ทุกสถานการณ์จะทำการทำซ้ำจำนวน 1,000 ครั้ง ซึ่งประสิทธิภาพของวิธีการตรวจสอบค่าผิดปกติจะพิจารณาจากร้อยละของความถูกต้องในการตรวจสอบ ผลการทดลองสรุปได้ดังนี้

ตัวสถิติ R_{2z}^2 จะเหมาะสำหรับใช้ในการตรวจสอบค่าผิดปกติเมื่อร้อยละของค่าผิดปกติค่อนข้างมาก คือตั้งแต่ร้อยละ 30 ขึ้นไป ที่ $p \geq 16$ และ $n \geq 50$ นั่นคือควรนำไปใช้กับข้อมูลที่มีจำนวนตัวแปรค่อนข้างมาก โดยมีจำนวน 16 ตัวแปรขึ้นไป และจำนวนข้อมูลตัวอย่างตั้งแต่ 50 ชุดข้อมูลขึ้นไป ส่วนตัวสถิติ R_{3z}^2 ยังคงให้ประสิทธิภาพที่ต่ำที่สุดในหลายกรณี จึงยังคงไม่เหมาะที่จะนำไปใช้ในการตรวจสอบค่าผิดปกติ ซึ่งอาจจะทำการศึกษาเพิ่มเติม เพื่อพัฒนาตัวสถิติดังกล่าวให้มีประสิทธิภาพสูงขึ้น สำหรับตัวสถิติ $d_{2i,(2)}^2$ จะเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อมีร้อยละของค่าผิดปกติตั้งแต่ 30 ขึ้นไป เมื่อมีจำนวนตัวแปรไม่มากเท่าใดนัก นั่นคือ $p \geq 10$ เช่นเดียวกับตัวสถิติ $d_{2i,(3)}^2$ จะเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อจำนวนตัวแปรตั้งแต่ 12 ถึง 14 ตัวแปร และตัวสถิติ d^2 จะเหมาะสำหรับนำไปใช้ตรวจสอบค่าผิดปกติเมื่อร้อยละของค่าผิดปกติไม่มากนัก และมีจำนวนตัวแปรไม่มากเช่นเดียวกัน แต่ในสถานการณ์จริงนักสถิติจะไม่สามารถทราบได้ชัดเจนว่าในชุดข้อมูลจริงเหล่านั้น จะมีค่าผิดปกติปะปนอยู่เป็นจำนวนมากน้อยเพียงใด ดังนั้นปัจจัยด้านร้อยละของค่าผิดปกติจึงไม่สามารถนำมาเป็นปัจจัยหลักในการกำหนด

เงื่อนไขของความเหมาะสมในการใช้ตัวสถิติเหล่านี้ เนื่องจากเป็นปัจจัยที่ไม่สามารถทราบได้ล่วงหน้า

สำหรับการนำวิธีการตรวจสอบค่าผิดปกติมาใช้กับข้อมูลด้านสินเชื่อนั้น ผลการตรวจสอบของตัวสถิติ R_{2z}^2 และ R_{3z}^2 จะได้ผลเช่นเดียวกัน ส่วนตัวสถิติ $d_{2i,(2)}^2$ และ $d_{2i,(3)}^2$ จะมีผลแตกต่างกันในบางชุดข้อมูล แต่ส่วนใหญ่จะเป็นชุดข้อมูลเดียวกัน สำหรับตัวสถิติ d^2 จะตรวจสอบค่าผิดปกติซึ่งจะพบค่าผิดปกติเป็นจำนวนมาก เนื่องจากเป็นผลจากการใช้ทุกองค์ประกอบมาสร้างตัวสถิติ จึงอาจทำให้ตัวสถิติ d^2 เกิดปัญหา swamping effect หรือ making effect ในบางชุดข้อมูล จึงควรทำการพิจารณาชุดข้อมูลนั้นๆ เพิ่มเติมในรายชุดข้อมูลต่อไป เพื่อให้ได้ข้อสรุปที่ถูกต้อง

ข้อเสนอแนะ

การวิจัยนี้นำเสนอตัวสถิติดังกล่าวเพื่อเป็นทางเลือกหนึ่งของตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติสำหรับข้อมูลหลายตัวแปร กรณีที่ค่าผิดปกติเกิดจากความสัมพันธ์ของตัวแปรไม่สอดคล้องกัน ซึ่งในปัจจุบันมีไม่มากนัก โดยค่าผิดปกติในลักษณะนี้พบมากกับข้อมูลทางการแพทย์ เศรษฐกิจ และนิเวศวิทยา สำหรับในการศึกษาวิจัยต่อไป ผู้วิจัยสามารถทำการศึกษาเพิ่มเติม ภายใต้ข้อสมมติของรูปแบบการแจกแจงของข้อมูลในแบบอื่น หรืออาจจะใช้แนวคิดของวิธีการทางสถิติอื่นๆ เช่น การวิเคราะห์ปัจจัย (factor analysis) และการวิเคราะห์กลุ่ม (cluster analysis) เป็นต้น ซึ่งเป็นวิธีการที่ใช้วิเคราะห์กับข้อมูลหลายตัวแปรมาปรับใช้เพื่อหาตัวสถิติทดสอบในการตรวจสอบหาค่าผิดปกติ รวมทั้งการพิจารณาจำนวนองค์ประกอบรองที่เหมาะสมต่อไป

เอกสารและสิ่งอ้างอิง

Bacon-Shone, J. and W.K. Fung. 1987. A new graphical method for detecting single and multiple outliers in univariate and multivariate data. **J. of the Royal Statistical Soc. Ser. C (Applied Stats.)** 36: 153-162.

Barnett, V. and T. Lewis. 1994. **Outlier in Statistical Data.** 3rd ed., John Wiley and Sons, Inc., New York.

Campbell, N.A. 1980. Robust procedures in multivariate analysis I: robust covariance estimation. **Applied Stats.** 29: 231-237.

Caroni, C. 2000. Outlier detection by robust principal components analysis. **Communications in Stats. – Simulation and Computation** 29: 139-151.

Croux, C. and G. Haesbroeck. 2000. Principal component analysis on robust estimators of the covariance or correlation matrix: inference functions and efficiencies. **Biometrika** 87: 603-618.

Davies, L. and U. Gather. 1993. The identification of multiple outliers. **J. of the Amer. Statistical Associ.** 88: 782-792.

Evans, M., N. Hastings and J.B. Peacock. 2000. **Statistical Distribution.** 3rd ed., John Wiley and Sons, Inc., New York.

Filzmoser, P., R. Maronna and M. Werner. 2008. Outlier identification in high dimensions. **Computational Stats. & Data Analysis** 52: 1694-1711.

Gnanadesikan, R. and J.R. Kettenring. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. **Biometrics** 28: 81-124.

Grubbs, F.E. 1969. Procedures for detecting outlying observations in samples. **Technometrics** 11: 1-21.

Hadi, A.S. 1992. Identifying outliers in multivariate data. **J. of the Royal Statistical Soc. Ser. B (Methodological)** 54: 761-771.

Hadi, A.S. 1994. A modification of a method for the detection of outliers in multivariate samples. **J. of the Royal Statistical Soc. Ser. B (Methodological)** 56: 393-396.

Hawkins, D.M. 1974. The detection of errors in multivariate data using principal components. **J. of the Amer. Statistical Associ.** 69: 340-344.

Hawkins, D.M. 1980. **Identification of Outliers.** Chapman and Hall, London.

Hawkins, D.M. and L.P. Fatti. 1984. Exploring multivariate data using the minor principal components. **J. of the Royal Statistical Soc. Ser. D (The Statistician)** 33: 325-337.

Jackson, D.A. and Y. Chen. 2004. Robust principal component analysis and outlier detection with ecological data. **Environmetrics** 15: 129-139.

Jobson, J.D. 1992. **Applied Multivariate Data Analysis.** Springer-Verlag Inc., New York.

Johnson, R.A. and D.W. Wichern. 2007. **Applied Multivariate Statistical Analysis.** 6th ed., Pearson Prentice Hall, New Jersey.

Jolliffe, I.T. 2002. **Principal Component Analysis.** 2nd ed., Springer-Verlag Inc., New York.

Kitagawa, G. 1979. On the use of AIC for the detection of outliers. **Technometrics** 21: 193-199.

Kosinski, A.S. 1999. A procedure for the detection of multivariate outliers. **Computational Stats. & Data Analysis** 29: 145-161.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. pp. 49-55. *In* **Proceedings the National Institute of Sciences of India.** 15 April 1936, Statistical Laboratory, Presidency College. Calcutta, India.

Marden, J.I. 1999. Some robust estimates of principal components. **Stats & Probability Letters** 43: 349-359.

Munoz-Garcia, J., J.L. Moreno-Rebollo and A. Pasual-Acosta. 1990. Outliers: a formal approach. **Int. Statistical Rev./Revue Internationale de Statistique** 58: 215-226.

Pena, D. and F.J. Prieto. 2001. Multivariate outlier detection and robust covariance matrix estimation. **Technometrics** 43: 286-300.

Penny, K.I. and Jolliffe, I.T. 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety data. **J. of the Royal Statistical Soc. Ser. D (The Statistician)** 50: 295-308.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. **Sankhya: The Indian J. of Stats., Ser. A** 26: 329-358.

Rencher, A.C. 1998. **Multivariate Statistical Inference and Applications.** John Wiley & Sons, Inc., New York.

Rencher, A.C. 2002. **Methods of Multivariate Analysis.** 2nd ed., John Wiley & Sons, Inc., New York.

Rocke, D.M. and D.L. Woodruff. 1996. Identification of outliers in multivariate data. **J. of the Amer. Statistical Associ.** 91: 1047-1061.

Rosner, B. 1975. On the detection of many outliers. **Technometrics** 17: 221-227.

Schwager, S.J. and B.H. Margolin. 1982. Detection of multivariate normal outliers. **The Ann. of Stats.** 10: 943-954.

Shan, Y. 2007. **Outlier Detection Using the Smallest Kernel Principal Components.** Ph.D. Dissertation, Temper University.

Shyu, M.L., S.C. Chen, K. Sarinnapakorn and L.W. Chang. 2003. A novel anomaly detection scheme based on principal component classifier, pp. 172-179. *In* **Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop.** 19 November 2003, The Third IEEE International Conference on Data Mining (ICDM'03). Melbourne, Florida, USA.

Wilk, S.S. 1963. Multivariate statistical outliers. **Sankhya: The Indian J. of Stats. Ser. A** 25: 407-426.

ภาคผนวก

ข้อมูลจริงที่ใช้ในการศึกษาครั้งนี้ประกอบด้วยตัวแปรอัตราส่วนทางการเงินมีจำนวน 10 ตัวแปร และจำนวนข้อมูลตัวอย่าง 65 ชุดข้อมูล โดยที่

x_1 คือ อัตราการหมุนเวียนของสินทรัพย์รวม (เท่า)

เป็นค่าที่แสดงถึงประสิทธิภาพในการใช้สินทรัพย์ทั้งหมดเมื่อเทียบกับยอดขาย โดยถ้ามีค่าน้อยจะแสดงว่า บริษัทมีสินทรัพย์มากเกินความต้องการ

x_2 คือ อัตราส่วนทุนหมุนเวียนเร็ว

เป็นค่าที่แสดงถึงส่วนของสินทรัพย์ระยะสั้น และมีความคล่องตัวในการเปลี่ยนเป็นเงินสด สามารถใช้แสดงถึงสภาพคล่องที่แท้จริงของกิจการได้ โดยปกติถ้ามีค่าเท่ากับ 1 จะถือว่าทุนหมุนเวียนเร็วมีความเหมาะสม

x_3 คือ อัตราส่วนสภาพคล่อง

เป็นค่าที่แสดงถึงส่วนของความสามารถในการชำระหนี้ระยะสั้น โดยถ้ามีค่ามากจะแสดงว่า บริษัทมีสินทรัพย์หมุนเวียนที่ประกอบไปด้วยเงินสด ลูกหนี้ และสินค้าคงเหลือมากกว่าหนี้ระยะสั้น ทำให้สภาพคล่องในการชำระหนี้ระยะสั้นมีค่อนข้างมาก โดยปกติถ้ามีค่าเท่ากับ 2 จะถือว่ามีสภาพคล่องเหมาะสม

x_4 คือ อัตราส่วนหนี้สินต่อทุน

เป็นค่าที่แสดงถึงความเสี่ยงในด้านเจ้าหนี้และเจ้าของกิจการ โดยถ้ามีค่ามากจะแสดงว่ากิจการมีความเสี่ยงจากการกู้ยืมเงินมาใช้ในการดำเนินกิจการ

x_5 คือ อัตรากำไรขั้นต้น (ร้อยละ)

เป็นค่าที่แสดงถึงประสิทธิภาพในการดำเนินงานของบริษัทในการทำกำไร ภายหลังหักต้นทุนเพียงอย่างเดียว

x_6 คือ อัตรากำไรสุทธิ (ร้อยละ)

เป็นค่าที่แสดงถึงประสิทธิภาพในการดำเนินงานของบริษัทในการทำกำไร ภายหลังหักต้นทุน ค่าใช้จ่ายรวมทั้งภาษีเงินได้หมดแล้ว

x_7 คือ อัตราผลตอบแทนจากสินทรัพย์รวม (ร้อยละ)

เป็นค่าที่แสดงถึงการวัดความสามารถในการทำกำไรของสินทรัพย์ทั้งหมดที่ใช้ในการดำเนินงานว่าให้ผลตอบแทนจากการดำเนินงานได้มากน้อยเพียงใด โดยถ้ามีค่ามากจะแสดงว่าบริษัทมีการใช้สินทรัพย์อย่างมีประสิทธิภาพ

x_8 คือ อัตราผลตอบแทนผู้ถือหุ้น (ร้อยละ)

เป็นค่าที่แสดงถึงเงินลงทุนในส่วนของเจ้าของ จะได้รับผลตอบแทนกลับคืนมาจากการดำเนินการของกิจการนั้นในอัตราส่วนเท่าไร โดยถ้ามีค่ามากจะแสดงว่า บริษัทมีประสิทธิภาพในการหากำไรสูง

x_9 คือ อัตราหมุนเวียนของสินค้าคงเหลือ

เป็นค่าที่แสดงถึงความสามารถในการบริหารการขายสินค้า โดยถ้ามีค่ามากจะแสดงว่าบริษัทมีความสามารถในการขายสินค้าได้เร็ว

x_{10} คือ อัตราหมุนเวียนของสินทรัพย์ถาวร (เท่า)

เป็นค่าที่แสดงถึงสินทรัพย์ถาวรทั้งหมดเมื่อเทียบกับยอดขาย โดยถ้ามีค่ามากจะแสดงว่าบริษัทมีการหมุนเวียนของสินทรัพย์ถาวรดี

โดยค่าของตัวแปรของข้อมูลสินเชื่อทั้งหมด แสดงดังตารางผนวกที่ 1

ตารางผนวกที่ 1 แสดงข้อมูลสินเชื่อเพื่อการลงทุน 10 ตัวแปร จำนวน 65 ชุดข้อมูล

ชุดข้อมูลที่	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	1.57393	0.03284	0.97438	1.28319	0.41195	7.01429	11.03999	30.24024	2.15531	4.47687
2	0.34674	0.6648	0.68682	0.40028	0.45797	17.15053	5.94672	10.0103	251	0.55781
3	1.45726	0.54857	0.71606	1.94509	0.11647	8.38816	12.2237	62.08351	16.69766	3.44215
4	0.53002	0.26823	0.55298	0.81844	0.13267	6.89225	3.65303	11.27427	6.02329	1.12937
5	1.56149	0.99065	1.47157	3.6286	0.32434	1.18728	1.85394	9.41513	5.47219	9.87451
6	0.8868	1.73322	2.06721	1.58259	0.35313	4.94582	4.38593	8.9166	5.07352	4.00269
7	1.68535	0.50494	1.21277	1.30836	0.10524	1.20637	2.03316	4.96777	5.83174	3.65977
8	5.47646	0.92605	0.96755	6.83601	0.01559	0.20164	1.10428	8.90629	152.0735	35.93165
9	1.22552	0.53445	2.79471	2.88973	0.32481	1.20905	1.48172	5.42051	1.31465	5.83357
10	0.71372	0.07947	0.34396	5.684	0.17584	0.60283	0.43025	16.97746	6.43966	0.83383
11	1.33316	0.73574	1.24447	1.44963	0.25051	0.20749	0.27661	0.60937	3.89383	4.83541
12	1.15876	1.47013	1.66099	4.34597	0.0586	-2.34704	-2.71966	-18.1114	14.60274	3.33558
13	2.31463	0.17663	2.02523	1.46118	0.14724	-0.30963	-0.71668	-1.33919	4.21054	10.61561
14	0.91718	2.7048	3.25839	0.37657	0.20507	2.16888	1.98925	2.21915	251	25.91105
15	2.5108	0.96472	1.25291	3.46266	0.19598	3.2093	8.05792	30.49915	11.38708	54.34556
16	1.59844	0.88211	1.00651	5.05735	0.12773	4.46197	7.13219	43.53302	24.90921	13.03992
17	1.32702	1.69529	2.34946	1.18074	0.23712	4.34707	5.76863	9.88015	6.86517	4.27223
18	0.79774	0.30953	2.2328	1.54082	0.32094	3.79188	3.02496	5.11243	2.51032	9.0328

ตารางผนวกที่ 1 (ต่อ)

ชุดข้อมูลที่	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
19	5.56081	0.5053	0.77303	1.58923	0.07167	0.8475	4.71279	15.54313	78.30737	10.73237
20	1.19193	0.85056	1.43209	0.81114	0.26887	9.87605	11.77155	19.61453	4.595	2.32255
21	1.78386	0.39474	0.97323	4.58477	0.20134	3.02957	5.40433	30.86345	3.08009	15.88653
22	1.28035	2.39996	2.55666	0.86854	-0.29783	8.03144	10.28307	13.7764	52.30286	3.64676
23	1.91793	0.26974	0.818	1.81131	0.20871	3.00194	5.7575	20.64893	5.09199	3.89548
24	1.26128	0.46784	0.77796	0.83771	0.05251	0.78767	0.99347	2.36384	8.71566	2.4274
25	3.37596	0.39283	1.24226	4.53128	0.07682	0.43543	1.47001	7.34821	5.22489	37.60982
26	1.51761	0.91874	1.09938	2.1095	0.11894	2.29247	3.47909	10.15481	11.73323	5.47331
27	0.79861	1.03092	2.18009	0.86713	0.2109	6.21242	4.96128	8.1658	2.65429	1.70381
28	1.44078	2.96436	3.29864	1.29687	0.56722	14.41184	20.76433	28.92789	11.11111	20.84756
29	0.77307	1.11734	1.42194	1.51546	0.10029	2.65487	2.05239	7.531	92.22018	1.53464
30	1.77954	1.34309	1.49181	1.31725	0.39243	7.56949	13.4702	26.39426	19.8262	5.4353
31	0.66915	0.66799	1.47045	1.06557	0.58808	-0.09134	-0.06112	-0.10619	1.21069	10.48994
32	1.74137	0.35008	0.65056	1.16907	-0.00439	-12.8266	-22.3359	-67.5288	10.21403	2.83926
33	0.69825	0.53125	1.22959	0.60626	0.1659	10.11093	7.05992	13.67781	3.8253	1.01626
34	1.27296	0.52739	1.22661	0.83573	0.17119	5.10509	6.49855	11.05965	5.28269	2.55644
35	2.46563	0.50527	1.0426	0.84092	0.08242	2.41029	5.94289	10.78467	11.24845	4.59484
36	0.7394	0.24196	1.34504	2.81981	0.18528	2.49878	1.84759	20.17247	4.73746	0.99685

ตารางผนวกที่ 1 (ต่อ)

ชุดข้อมูลที่	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
37	2.0209	0.83279	1.31401	1.46885	0.11387	3.56362	7.20171	15.25203	12.33245	6.59771
38	0.39738	0.30881	2.38067	1.10008	0.32727	20.09569	7.98557	20.24679	0.75838	0.84605
39	1.85677	0.51194	0.89296	1.02198	0.10869	3.5532	6.59745	14.14812	16.76301	3.61963
40	1.84459	1.19434	1.34747	3.12101	0.21187	3.86079	7.12155	23.61647	31.71996	31.33933
41	3.59285	0.16866	0.71503	1.89953	0.16934	1.99262	7.15917	27.16906	9.88633	7.19339
42	2.27976	1.06761	1.07291	5.69603	0.10299	3.4913	7.95934	52.53886	251	16.69304
43	2.45483	0.49325	1.08057	0.81579	0.15635	8.2007	20.1313	38.65887	36.22998	4.26791
44	1.00414	0.80199	1.18184	5.38684	0.10306	1.37329	1.37898	7.66437	251	37.36911
45	0.06607	0.95924	0.9698	0.32861	0.76921	8.13493	0.53747	3.74616	251	0.06934
46	1.11736	0.80201	2.38912	1.03395	0.19505	14.76885	16.50217	38.64036	5.75096	2.0009
47	1.33289	0.35177	1.32836	3.11762	0.14108	1.59029	2.11968	7.96811	1.95183	7.81069
48	0.53625	0.15177	1.36968	2.36052	0.15491	4.85336	2.60259	7.08792	13.96492	4.02441
49	0.82091	0.47077	0.96093	0.53256	0.16727	10.3731	8.51543	16.48421	5.52654	2.02405
50	1.68572	1.07626	2.41289	0.77013	0.05964	4.70122	7.92493	21.1637	13.08008	2.59096
51	1.57525	0.42058	1.42668	1.16153	0.15781	3.68232	5.80058	10.69926	3.25346	4.63524
52	3.60295	0.429	2.15679	1.46868	0.20394	3.92741	14.15029	25.01238	9.94601	24.58966
53	2.68689	1.46521	2.0721	1.74056	0.281	12.07658	32.44848	59.70521	251	50.49787
54	0.93828	0.24122	0.44634	0.69247	0.52729	1.72573	1.61922	10.56014	10.80183	42.65698

ตารางผนวกที่ 1 (ต่อ)

ชุดข้อมูลที่	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
55	1.58026	1.33538	2.02576	1.25472	0.15097	3.56534	5.63416	9.12385	251	16.53203
56	0.93383	0.60632	0.99728	0.88846	1	0.40845	0.38142	1.22289	0	1.57397
57	4.22237	0.24569	1.52839	1.8372	0.28602	4.18116	17.65442	38.87594	4.3529	59.53419
58	2.04862	1.00837	1.29093	3.93186	0.17579	3.19503	6.54541	26.48111	251	72.77329
59	4.21951	0.83144	1.68542	0.79559	0.17404	3.88655	16.39935	24.76133	39.48534	39.59107
60	1.8469	0.71339	1.06807	3.708	0.28594	2.30523	4.25755	19.03837	9.62365	18.86751
61	1.20192	1.70816	1.93696	1.58603	1	18.71421	22.49293	42.17167	0	9.27982
62	2.53126	0.98826	1.0401	7.33571	0.0994	2.3567	5.96541	48.03894	50.21485	28.4217
63	1.83805	0.65013	2.04424	1.82537	0.25859	0.6168	1.13371	2.93414	2.8651	6.23692
64	1.68529	0.23716	0.56824	1.56371	0.0851	0.32993	0.55604	2.3604	7.85243	3.03133
65	1.95704	0.36942	1.20612	2.92642	0.06765	1.51041	2.95593	10.84478	88.44879	11.74582

ประวัติการศึกษาและการทำงาน

ชื่อ	นางรุ้งรวี อำนาจตระกูล
เกิดวันที่	29 มีนาคม 2521
สถานที่เกิด	อำเภอเมือง จังหวัดฉะเชิงเทรา
ประวัติการศึกษา	วท.บ. (คณิตศาสตร์) มหาวิทยาลัยธรรมศาสตร์
	วท.ม. (สถิติ) มหาวิทยาลัยเกษตรศาสตร์
ตำแหน่งปัจจุบัน	อาจารย์ (พนักงานในสถาบันอุดมศึกษา สายวิชาการ)
สถานที่ทำงานปัจจุบัน	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏราชนครินทร์
ทุนการศึกษาที่ได้รับ	ได้รับทุนพัฒนาอาจารย์จากมหาวิทยาลัยราชภัฏราชนครินทร์ (พ.ศ. 2549)