การวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพงานในกระบวนการทดสอบฮาร์ดดิสก์

นางสาวพฤกษารัตน์ สิทธิพงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิศวกรรมระบบอุตสาหกรรมและสิ่งแวดล้อม

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2560

RELATIONSHIPS ANALYSIS FOR PROCESS VARIABLES THAT AFFECT HARD DISK QUALITY

Preuksarat Sittipong

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Industrial Engineering

Suranaree University of Technology

Academic Year 2017

การวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพงานใน กระบวนการทดสอบฮาร์ดดิสก์

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

(รศ. ดร.พรศิริ จงกล)
ประธานกรรมการ

(อ. ดร.นรา สมัตถภาพงศ์)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

(รศ. ดร.นิวิท เจริญใจ)
กรรมการ

(ผศ. ดร.ปวีร์ ศิริรักษ์)
กรรมการ

(ศ. ดร.สันติ แม้นศิริ)
รองอธิการบดีฝ่ายวิชาการและพัฒนาความเป็นสากล

(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

พฤกษารัตน์ สิทธิพงศ์ : การวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพ งานในกระบวนการทดสอบฮาร์ดดิสก์ (RELATIONSHIPS ANALYSIS FOR PROCESS VARIABLES THAT AFFECT HARD DISK QUALITY) อาจารย์ที่ปรึกษา : อาจารย์ ดร.นรา สมัตถภาพงศ์, 73 หน้า.

การสูญเสียของกระบวนการผลิตฮาร์ดดิสก์ สาเหตุมาจากผลิตภัณฑ์ฮาร์ดดิสก์บางส่วนไม่ผ่านกระบวนการตรวจสอบ โดยขั้นตอนกระบวนการตรวจสอบของคุณภาพฮาร์ดดิสก์มีความซับซ้อนมากทำให้ยากต่อการตรวจสอบคุณภาพ กระบวนการผลิตฮาร์ดดิสก์ในขั้นตอนการตรวจสอบคุณภาพนั้นจะมีการวัดค่าตัวแปรจากกระบวนการต่าง ๆ ในการผลิต ซึ่งตัวแปรเหล่านี้มีผลกับคุณภาพของฮาร์ดดิสก์ที่จะผ่านขั้นตอนการตรวจสอบคุณภาพหรือไม่ ด้วยตัวแปรในกระบวนการผลิตที่หลากหลาย ซึ่งมีทั้งที่เป็นตัวแปรเชิงปริมาณ (Quantitative Variable) และตัวแปรเชิงคุณภาพ (Qualitative Variable) หากผู้ผลิตสามารถทราบถึงตัวแปรหลักที่มีผลกระทบต่อคุณภาพของฮาร์ดดิสก์ ในการผลิตแต่ละรอบแล้ว จะสามารถวิเคราะห์ปัญหา และหาสาเหตุที่ทำให้ฮาร์ดดิสก์ไม่ผ่านการทดสอบได้ อีกทั้งยังส่งผลต่อการเพิ่มปริมาณฮาร์ดดิสก์ที่ผ่านการทดสอบได้มากขึ้น แต่เนื่องด้วยปริมาณของตัวแปรจำนวนมากและข้อมูลที่มหาศาล จึงทำให้ประสบปัญหาในการเลือกใช้เครื่องมือและ วิธีการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต่าง ๆ และผลของการทดสอบ

การศึกษาวิจัยนี้มีวัตถุประสงค์คือ เพื่อหาค่าความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ที่มีผลต่อการทดสอบคุณภาพของฮาร์ดดิสก์และ เพื่อศึกษาตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ของบริษัทกรณีศึกษา โดยข้อมูลที่ได้จากบริษัทกรณีศึกษาเป็นค่าที่วัดจากกระบวนการตรวจสอบคุณภาพของฮาร์ดดิสก์ ข้อมูลที่ได้รับมีจำนวนตัวแปรคุณภาพ (ตัวแปรตาม) ของฮาร์ดดิสก์ทั้งหมด 57,232 ตัว และจำนวนตัวแปรที่เกี่ยวข้องต่อคุณภาพของฮาร์ดดิสก์ (ตัวแปรอิสระ) 74 ตัว ซึ่งแบ่งเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจำนวน 53 ตัวและ ตัวแปรเชิงปริมาณจำนวน 21 ตัว การวิจัยนี้ใช้การหาค่าความสัมพันธ์ของข้อมูลด้วยวิธี Mutual Information โดยใช้โปรแกรม R และทวนสอบผลด้วยวิธี Logistic Regression โดยใช้โปรแกรม SPSS®

ผลที่ได้จากการศึกษาพบว่า การวิเคราะห์ความสัมพันธ์ของตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์จากทั้ง 2 วิธีสามารถสรุปตัวแปรอิสระที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์เหมือนกันดังต่อไปนี้ ตัวแปร A_07, A_18, A_19, A_20, A_21 และ A_28

สาขาวิชา วิศวกรรมอุตสาหการ ลายมือชื่อนักศึกษา พฤกษารัตน์

ปีการศึกษา 2560 ลายมือชื่ออาจารย์ที่ปรึกษา

PREUKSARAT SITTIPONG : RELATIONSHIPS ANALYSIS FOR PROCESS VARIABLES THAT AFFECT HARD DISK QUALITY. THESIS ADVISOR : NARA SAMATTAPAPONG, Ph.D., 73 PP.

HARD DISK/FAILURE/QUALITY/CORRELATION

Losing of hard disk manufacturing process, the caused was from the hard disk's products did not past the verification process. The verification process of hard disk was very complicated and very hard for verification of quality. In quality checking process of hard disk, manufacturing were measured by many variables and variety of processes. These variables affected directly in the quality checking process. In variety of manufacturing, there had many variables which included "Quantitative Variable" and "Qualitative Variable". If the manufacturers know about how the main variable that affected to hard disk's quality when they manufacture in each time, the issues and cause that made hard disk did not pass the quality test would be analyzed, also affect to the increasing the amount of hard disk, it would give more chance to pass the quality. Since there is a huge amount of variables and enormous information, it is hard to use the tools and analyze the relation of variables and the result of the test.

The purposes of this research were to find out the relation value of variables that affected to the quality test of hard disk, and to study how the variables affected to the quality of hard disk from the company's case study. The information that the researcher got from the company was a value measurement from hard disk's quality checking process, there were 57,232 variables from qualitative variable (dependent variables) of hard disk, and the variable that related to quality of hard disk

(independent variables) 74 variables. The researcher has classified 53 variables of qualitative variable, and 21 variables of quantitative variable. This research was used to find the value relation of information with Mutual Information method by used program R, and verified with Logistic Regression method by used program SPSS®.

The result of study was found that analyzing the relation of variable which affected to quality to hard disk by 2 methods can summarized that, the independent variable that affected to quality of hard disk were same as follows; variable A_07, A_18, A_19, A_20, A_21, and A_28.

School of Industrial engineering Student's Signature พฤกษารัตน์

Academic Year 2017 Advisor's Signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี เนื่องจากได้รับความช่วยเหลืออย่างดียิ่ง ทั้งด้านวิชาการ ด้านการดำเนินงาน คำปรึกษา แนะนำ ช่วยเหลือ และด้านอื่น ๆ ที่เกี่ยวข้อง จากบุคคลและกลุ่มบุคคลต่าง ๆ ได้แก่

ขอกราบขอบพระคุณ อาจารย์ ดร.นรา สมัตถภาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำปรึกษามาโดยตลอด ทั้งด้านวิชาการ คำแนะนำ รวมทั้งช่วยตรวจทานแก้ไขวิทยานิพนธ์เล่มนี้จนเสร็จสมบูรณ์

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.พรศิริ จงกล ประธานกรรมการสอบโครงร่างวิทยานิพนธ์ ที่ให้คำเสนอแนะในการปรับปรุงแก้ไขวิทยานิพนธ์

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.นิวิท เจริญใจ อาจารย์ภาควิชาวิศวกรรมอุตสาหการ มหาวิทยาลัยเชียงใหม่ สำหรับการให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์

ขอขอบพระคุณ คุณพงศกร สรณาคมน์ คุณจารุพงศ์ สุทธมงคล และคุณศุภกาญจน์ ศุภสุข นักศึกษาระดับปริญญาตรี สาขาวิศวกรรมอุตสาหการ มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ร่วมมือทำงานสำเร็จลุล่วงด้วยดี

ขอขอบพระคุณ มหาวิทยาลัยเทคโนโลยีสุรนารีที่ให้โอกาสทางการศึกษา

สุดท้ายนี้ ผู้วิจัยขอขอบพระคุณ บิดา มารดา และครอบครัว ซึ่งเป็นกำลังใจอันสำคัญและ เปิดโอกาสให้ได้รับการศึกษาเล่าเรียน คอยช่วยเหลือ ที่ทำให้วิทยานิพนธ์สำเร็จได้ด้วยดี

พฤกษารัตน์ สิทธิพงค์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ประเทศไทยนับเป็นฐานการผลิตฮาร์ดดิสก์ที่สำคัญแห่งหนึ่งของโลก เป็นอันดับสองรองจากจีน (ศูนย์วิจัยกสิกรไทย, 2558) โดยอุตสาหกรรมคอมพิวเตอร์และ ส่วนประกอบของประเทศไทยมีฮาร์ดดิสก์เป็นผลิตภัณฑ์หลักที่ติดอันดับ 1 ใน 3 ของสินค้าที่สร้างรายได้ให้กับประเทศไทย (ฌาตยา แดงรุ่งโรจน์, 2014)

ฮาร์ดดิสก์ (Hard Disk) เป็นส่วนประกอบสำคัญของระบบคอมพิวเตอร์ ซึ่งเป็นแหล่งจัดเก็บข้อมูลของระบบ ถือว่ามีความสำคัญมากถ้าหากไม่มีฮาร์ดดิสก์คอมพิวเตอร์ก็จะไม่สามารถทำงานได้ เนื่องจากขาดสื่อหลักที่ใช้ในการเก็บบันทึกข้อมูลของเครื่องคอมพิวเตอร์ ด้วยการพัฒนาอย่างรวดเร็วของเทคโนโลยีคอมพิวเตอร์และ อินเทอร์เน็ตจำนวนข้อมูลจะเพิ่มขึ้นอย่างมาก คาดว่ากว่า 90% ของข้อมูลใหม่ทั้งหมดที่ผลิตในโลกจะถูกเก็บไว้ในสื่อแม่เหล็กส่วนใหญ่จะเป็นฮาร์ดดิสก์ (Pinheiro et al., 2007) โดยผู้ใช้งานทุกคนต่างต้องการความเชื่อมั่นในคุณภาพของฮาร์ดดิสก์ เนื่องจากผู้ใช้งานใช้ในการเก็บข้อมูลที่สำคัญ

ในปัจจุบันอุตสาหกรรมฮาร์ดดิสก์มีการแข่งขันสูง ผู้ผลิตฮาร์ดไดรฟ์ได้ทำการพัฒนาเทคโนโลยีการตรวจสอบผลิตภัณฑ์ตั้งแต่ปี 1994 (Joseph F. Murray et al., 2005) บริษัทกรณีศึกษาจึงต้องการเพิ่มประสิทธิภาพกระบวนการผลิตฮาร์ดดิสก์ การสูญเสียของกระบวนการผลิตฮาร์ดดิสก์สาเหตุมาจากผลิตภัณฑ์ฮาร์ดดิสก์บางส่วนไม่ผ่านกระบวนการตรวจสอบ (ธีรนพ สุขอารมณ์และ ปารเมศ ชุติมา, 2556) และด้วยกระบวนการผลิตของฮาร์ดดิสก์มีความซับซ้อนจึงทำให้ยากต่อการตรวจสอบคุณภาพ โดยกระบวนการผลิตฮาร์ดดิสก์ในขั้นตอนการตรวจสอบคุณภาพจะมีการวัดค่าตัวแปรจากกระบวนการต่าง ๆ ในการผลิต โดยที่กระบวนการคือ เครื่องทดสอบแต่ละเครื่องจะมีหุ่นยนต์หยิบชิ้นงานฮาร์ดดิสก์ที่ต้องการจะทดสอบเข้าเครื่องเพื่อทำการทดสอบและ เมื่อทดสอบเสร็จจะมีทั้งงานที่ผ่านการทดสอบ (PASS) หรือไม่ผ่านการทดสอบ (FAIL) (ประสาน นาคอ่อน, 2557) ซึ่งกระบวนการนี้จะให้ค่าตัวแปรต่าง ๆ ที่ส่งผลกับคุณภาพของฮาร์ดดิสก์ ที่จะผ่านขั้นตอนการตรวจสอบคุณภาพหรือไม่ โดยมีทั้งที่เป็นตัวแปรเชิงปริมาณ (Quantitative Variable) และตัวแปรเชิงคุณภาพ (Qualitative Variable) ทั้งนี้หากผู้ผลิตสามารถทราบถึงตัวแปรหลักที่มีผลกระทบต่อคุณภาพของฮาร์ดดิสก์ในการผลิตแต่ละรอบแล้ว จะสามารถวิเคราะห์ปัญหาและ หาสาเหตุที่ทำให้ฮาร์ดดิสก์ไม่ผ่านการทดสอบได้ ซึ่งจะนำไปสู่การปรับปรุงคุณภาพของการผลิตได้ถูกต้องและ

ส่งผลต่อการเพิ่มปริมาณฮาร์ดดิสก์ที่ผ่านการทดสอบได้มากขึ้น แต่เนื่องด้วยปริมาณของตัวแปรจำนวนมากและ ข้อมูลที่มหาศาล จึงทำให้ประสบปัญหาในการเลือกใช้เครื่องมือและ วิธีการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต่าง ๆ และผลของการทดสอบ

งานวิจัยที่เกี่ยวข้องกับการศึกษาความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อคุณภาพของฮาร์ดดิสก์ ภายในประเทศไทยยังมีอยู่บ้าง แต่ไม่มีการใช้ด้วยวิธีโปรแกรม R ในการหาค่าความสัมพันธ์ของข้อมูลด้านอุตสาหกรรมฮาร์ดดิสก์ โดยโปรแกรม R เป็น Open Source Software ที่มีลิขสิทธิ์แต่ไม่เสียค่าใช้จ่ายในการใช้งาน สามารถดาวน์โหลดมาใช้ได้ทุกที่ทั่วโลก (กาญจน์ คุ้มทรัพย์, 2559) มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูล สามารถดัดแปลงได้ตามต้องการ ใช้ได้ฟรีบนเครื่องคอมพิวเตอร์แบบต่าง ๆ ไม่ว่าจะใช้บน Windows, Mac OS หรือ Linux ภาษา R พัฒนามาจากภาษา S โดยพัฒนาขึ้นมาเพื่อใช้ในงานสถิติ (วิโรจน์ อรุณมานะกุล, 2559) เนื่องจากโปรแกรม R เป็นโปรแกรมฟรี ซึ่งช่วยลดต้นทุนของอุตสาหกรรมฮาร์ดดิสก์ในด้านการวิเคราะห์ข้อมูลได้ จึงเป็นโปรแกรมที่น่าสนใจในการนำมาใช้งาน

จากปัญหาดังกล่าวทางผู้วิจัยจึงได้เล็งเห็นความสำคัญของการใช้โปรแกรม R เข้ามาช่วยในการศึกษาความสัมพันธ์ของข้อมูลที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์และ เพื่อหาแนวทางในการแก้ไขปัญหาได้ตรงสาเหตุมากขึ้น จนในที่สุดก็จะเพิ่มประสิทธิภาพกระบวนการผลิตฮาร์ดดิสก์ของบริษัทกรณีศึกษา

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อหาค่าความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ที่มีผลต่อการทดสอบคุณภาพของฮาร์ดดิสก์

1.2.2 เพื่อศึกษาตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ ของบริษัทกรณีศึกษา

1.3 ขอบเขตการวิจัย

1.3.1 ศึกษาข้อมูลตัวอย่างของฮาร์ดดิสก์ทั้งหมด 57,232 ตัวอย่าง จากตัวแปร 74 ตัว แบ่งเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจำนวน 53 ตัวและ ตัวแปรเชิงปริมาณจำนวน 21 ตัว ซึ่งเป็นตัวแปรที่ได้จากกระบวนการทดสอบคุณภาพของฮาร์ดดิสก์เท่านั้น

1.3.2 ใช้โปรแกรม R เป็นเครื่องมือในการวิเคราะห์ความสัมพันธ์ของข้อมูล

1.3.3 ใช้เทคนิคการหาค่าความสัมพันธ์ด้วยวิธี Mutual Information

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 บริษัทอุตสาหกรรมฮาร์ดดิสก์สามารถนำงานวิจัยนี้ไปประยุกต์ใช้

1.4.2 ได้วิธีการที่ง่ายต่อการใช้งานสำหรับการศึกษาความสัมพันธ์ของตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์

1.4.3 เพื่อเป็นแนวทางในการพยากรณ์คุณภาพของฮาร์ดดิสก์ต่อไป

บทที่ 2
ทฤษฎีที่เกี่ยวข้องและปริทัศน์วรรณกรรม

การวิเคราะห์ความสัมพันธ์ของตัวแปรต่าง ๆ ที่มีผลในกระบวนการทดสอบคุณภาพของฮาร์ดดิสก์" มีจุดประสงค์หลักเพื่อหาค่าความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ที่มีผลต่อการทดสอบคุณภาพของฮาร์ดดิสก์และ เพื่อศึกษาตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ ในบทนี้จะกล่าวโดยสังเขปถึงทฤษฎีและ ปริทัศน์วรรณกรรมที่เกี่ยวข้องกับการวิเคราะห์ค่าความสัมพันธ์ของข้อมูล

2.1 ความหมายของ "สถิติ"

ความหมายของ "สถิติ" มีอยู่ 2 ประการ คือ

- หมายถึง ตัวเลขสถิติ ที่แสดงข้อเท็จจริงของสิ่งต่าง ๆ เช่น สิ่งของ บุคคล หรือปรากฏการณ์ต่าง ๆ โดยจะประกอบด้วยตัวเลขหลาย ๆ ตัว
- หมายถึง "สถิติศาสตร์" เป็นแขนงหนึ่งของวิธีการทางวิทยาศาสตร์ โดยประกอบด้วย การเก็บข้อมูลจากการสำรวจหรือทดลอง การจัดระบบ การนำเสนอข้อมูล การวิเคราะห์ข้อมูลและ การแปลความหมายของข้อมูล วัตถุประสงค์ของวิชาสถิติคือ บรรยายและสรุปคุณลักษณะของประชากร

(ชูศรี วงศ์รัตนะ, 2560)

ซึ่งวิชาสถิติศาสตร์แบ่งเป็น 2 กลุ่ม คือ

- สถิติเชิงอนุมาน (Inferential Statistics)
- สถิติเชิงพรรณา (Descriptive Statistics)

2.2 สถิติเชิงอนุมาน (Inferential Statistics)

สถิติเชิงอนุมาน (Inferential Statistics) เป็นการศึกษากลุ่มตัวอย่าง โดยใช้ทฤษฎีความน่าจะเป็นสรุปผลที่ศึกษาได้จากกลุ่มตัวอย่างและกลุ่มประชากรนั้น โดยกลุ่มตัวอย่างต้องเป็นข้อมูลตัวแทนที่ดีของกลุ่มประชากร

2.3 สถิติเชิงพรรณา (Descriptive Statistics)

สถิติเชิงพรรณา (Descriptive Statistics) ใช้ในการสรุปลักษณะของข้อมูลซึ่งจะได้จำนวน ค่าพิสัย ร้อยละ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน เป็นต้น จากข้อมูลที่ทำการเก็บรวบรวม (กัลยา วานิชย์บัญชา, 2549)

การเลือกใช้สถิติเชิงพรรณาในงานวิจัยแบ่งเป็น 2 กลุ่ม คือ สำหรับข้อมูลเชิงคุณภาพหรือข้อมูลเชิงกลุ่มและ สำหรับข้อมูลเชิงปริมาณซึ่งขึ้นกับสเกลหรือประเภทของข้อมูล อาจประกอบด้วยข้อมูลหลากหลายประเภททั้งข้อมูลเชิงปริมาณและหรือเชิงคุณภาพ หรือบางทีอาจประกอบด้วยข้อมูลทั้ง 4 สเกล

2.3.1 ประเภทของข้อมูล

ในการทำวิจัยควรศึกษาประเภทของข้อมูล เนื่องจากประเภทของข้อมูลจะกำหนดเทคนิคการวิเคราะห์ทางสถิติที่เราจะนำมาใช้ โดยการแบ่งประเภทของข้อมูลแบ่งได้หลายลักษณะดังนี้

- การแบ่งตามที่มาของข้อมูลสามารถแบ่งได้ 2 ประเภท คือ ข้อมูลปฐมภูมิ (Primary Data) และข้อมูลทุติยภูมิ (Secondary Data) โดยข้อมูลปฐมภูมิจะได้จากการที่นักวิจัยเก็บรวบรวมจากแหล่งข้อมูลเอง อาจใช้วิธีการทดลอง สัมภาษณ์ สังเกต เป็นต้น ส่วนข้อมูลทุติยภูมิเป็นข้อมูลที่ได้จากหน่วยงานหรือมีคนอื่นเก็บข้อมูลไว้แล้วซึ่งสามารถนำมาใช้ได้เลย ทำให้สะดวกแก่นักวิจัย แต่ข้อมูลที่มีอาจไม่เพียงพอ
- แบ่งตามลักษณะของข้อมูลซึ่งแบ่งได้เป็น ข้อมูลเชิงคุณภาพ (Qualitative Data) หรือเรียกอีกอย่างหนึ่งว่าข้อมูลเชิงกลุ่มและ ข้อมูลเชิงปริมาณ (Quantitative Data) โดยข้อมูลเชิงคุณภาพจะอยู่ในรูปของตัวแปรที่มีค่าได้ 2 ค่า เช่น “ผ่าน” กับ “ไม่ผ่าน” ส่วนข้อมูลเชิงปริมาณจะอยู่ในรูปที่เป็นตัวเลขที่สามารถวัดค่าได้ บ่งบอกความมากน้อยของตัวแปรได้
- แบ่งตามมาตรของข้อมูล แบ่งได้เป็น 4 มาตร ดังนี้ (Steven, 1960)

— มาตรนามบัญญัติ (Nominal Scale) เป็นการแบ่งข้อมูลออกเป็น 2 ประเภท ค่าตัวแปรเป็นการจำแนกประเภท เช่น “ผ่าน” กับ “ไม่ผ่าน” หรือจำแนกศาสนาที่คนนับถือ เช่น พุทธ คริสต์ อิสลาม เป็นต้น ไม่สามารถเรียงลำดับ หรือบอกปริมาณความแตกต่างได้

— มาตรอัตราส่วน (Ratio Scale) เป็นข้อมูลที่บอกความแตกต่างได้ดี สามารถเปรียบเทียบความแตกต่างได้ ค่าตัวแปรสามารถเรียงลำดับและบอกปริมาณความแตกต่างระหว่างแต่ละค่าได้อย่างชัดเจนและ มีค่าเป็นศูนย์แท้เช่น นาย A สูง 150 เซนติเมตร กับ นาย B สูง 300 เซนติเมตร ซึ่งบอกได้ว่านาย B สูงกว่านาย A 150 เซนติเมตรและ นาย B สูงกว่า นาย A สองเท่า

– มาตรอันตรภาค (Interval Scale) เป็นข้อมูลที่บอกความแตกต่างได้ บ่งบอกถึงปริมาณว่ามีมากกว่า น้อยกว่าเพียงใด ค่าตัวแปรสามารถเรียงลำดับและ บอกปริมาณความแตกต่างระหว่างแต่ละค่าได้อย่างชัดเจน ข้อมูลของมาตรอันตรภาคนี้จะไม่มีศูนย์แท้

– มาตรอันดับ (Ordinal Scale) ค่าตัวแปรเรียงลำดับได้ แต่ไม่สามารถบอกปริมาณความแตกต่าง ใช้ในการแบ่งกลุ่มแต่ไม่สามารถบอกปริมาณความแตกต่างระหว่างแต่ละค่าได้อย่างชัดเจน การจัดลำดับต้องมีเกณฑ์ช่วย เช่น ระดับการศึกษา ตำแหน่งทางการบริหาร

จากข้อมูล 4 มาตรที่ได้อธิบายข้างต้น สามารถแยกข้อมูลได้เป็น 2 กลุ่ม คือ

- ข้อมูลเชิงปริมาณ เป็นข้อมูลมาตรอัตราส่วน (Ratio Scale) และมาตรอันตรภาค (Interval Scale)
- ข้อมูลเชิงคุณภาพ เป็นข้อมูลมาตรนามบัญญัติ (Nominal Scale) และมาตรอันดับ (Ordinal Scale)

(ศิริชัย พงษ์วิชัย, 2556)

2.3.2 สรุปสถิติเชิงพรรณาที่ใช้ในงานวิจัย

การเลือกใช้สถิติพรรณนาในการวิจัยกับชนิดของสเกลข้อมูล (กัลยา วานิชย์บัญชา, 2549) ได้สรุปไว้ดังนี้

ตารางที่ 2.1 สรุปสถิติเชิงพรรณาที่ใช้ในงานวิจัย

ชนิดของข้อมูล	สถิติเชิงพรรณา
1. สเกลแบ่งกลุ่ม (Nominal Scale) 1 ตัว	- ความถี่ ร้อยละ - ฐานนิยม (Mode)
2. สเกลอันดับ (Ordinal Scale) 1 ตัว	- ความถี่ ร้อยละ - ค่ามัธยฐาน ฐานนิยม
3. ตัวแปรเชิงปริมาณ (Interval or Ratio Scale) 1 ตัว	- ค่ากลาง (ค่าเฉลี่ย ค่ามัธยฐาน ฐานนิยม) - ค่ากระจาย (ค่าแปรปรวน พิสัย ค่าเบี่ยงเบนมาตรฐาน)
4. ตัวแปรเชิงกลุ่ม 2 ตัว	- ตารางไขว้ (Crosstab) แสดงจำนวนและร้อยละ
5. ตัวแปรเชิงปริมาณ 1 ตัว และตัวแปรเชิงกลุ่ม 1 ตัว	- ค่ากลางและค่าการกระจายของตัวแปรเชิงปริมาณแยกตามกลุ่มย่อยของตัวแปรเชิงกลุ่ม
6. ตัวแปรเชิงปริมาณ 2 ตัว	- การวิเคราะห์ความถดถอย - สัมประสิทธิ์สหสัมพันธ์

เทคนิคที่ใช้ในการวัดความสัมพันธ์ของตัวแปร 2 ตัว มีหลายเทคนิค (กัลยา วานิชย์บัญชา, 2549) ได้สรุปดังตารางที่ 2.2 - 2.3

ตารางที่ 2.2 สรุปเทคนิคทางสถิติที่วิเคราะห์ความสัมพันธ์ของตัวแปร 2 ตัว

ชนิดตัวแปรต้นหรือตัวแปรอิสระ	ชนิดของตัวแปรตาม	เทคนิคการวิเคราะห์ทางสถิติ
1. ตัวแปรเชิงกลุ่ม 1 ตัว ซึ่งแบ่งเป็น 2 กลุ่มย่อย	ตัวแปรเชิงปริมาณ 1 ตัว	t-test, Z-test
2. ตัวแปรเชิงกลุ่ม 1 ตัว ซึ่งแบ่งเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม	ตัวแปรเชิงปริมาณ 1 ตัว	F-test (1-WAY ANOVA)
3. ตัวแปรเชิงปริมาณ 1 ตัว	ตัวแปรเชิงปริมาณ 1 ตัว	การวิเคราะห์ความถดถอยอย่างง่าย (Simple Regression)

ตารางที่ 2.3 การวิเคราะห์ความสัมพันธ์ 2 ตัวแปรที่ไม่มีการแบ่งเป็นตัวแปรต้นและตัวแปรตาม

ชนิดของตัวแปร 2 ตัว	เทคนิคการวิเคราะห์ทางสถิติ
1. เชิงกลุ่มทั้ง 2 ตัว	การทดสอบไคสแควร์
2. เชิงปริมาณทั้ง 2 ตัว	สัมประสิทธิ์สหสัมพันธ์เพียร์สัน
3. ตัวแปรสเกลอันดับ 2 ตัว	- สัมประสิทธิ์สหสัมพันธ์สเปียร์แมน - แกมม่า (gamma)

2.4 การวิเคราะห์การถดถอยและสหสัมพันธ์ (Regression and Correlation Analysis)

การวิเคราะห์การถดถอยและสหสัมพันธ์เป็นเครื่องมือทางสถิติอย่างหนึ่งที่ใช้ในการวิเคราะห์ตัวแปรตั้งแต่ 2 ตัวขึ้นไปว่ามีความเกี่ยวพันกันในรูปแบบใดและ ความสัมพันธ์มีมากน้อยเพียงใด วัตถุประสงค์ในการศึกษาความสัมพันธ์ระหว่างตัวแปร เพื่อวิเคราะห์ตัวแปรตัวหนึ่งเมื่อทราบค่าของตัวแปรอื่น ๆ ที่เกี่ยวข้อง ซึ่งตัวแปรที่ศึกษาต้องมีเพียงตัวเดียวเท่านั้น โดยการวิเคราะห์ในลักษณะนี้เรียกว่า การวิเคราะห์การถดถอยและสหสัมพันธ์อย่างง่าย (Simple Regression and Correlation Analysis) หากกรณีที่ต้องการศึกษามี 2 ตัวแปรขึ้นไปจะเรียกว่า การวิเคราะห์การถดถอยและสหสัมพันธ์เชิงพหุ (Multiple Regression and Correlation Analysis) (สมจิต วัฒนาชยากุล, 2535)

2.4.1 เส้นถดถอยแบบเส้นตรง (Linear Regression)

สมการความสัมพันธ์แบบเส้นตรงคือ $Y=\alpha+\beta X+e$ โดยที่ α และ β เป็นตัวคงที่ และ e คือความคลาดเคลื่อนที่เกิดขึ้น ซึ่งมีค่าเฉลี่ยเป็นศูนย์และ ความแปรปรวนเป็น $\sigma^2_{Y.X}$ ดังนั้น $\mu_{Y.X}= \alpha+\beta X$ หรือ $Y_e=\alpha+\beta$

ข้อตกลงเบื้องต้นในการวิเคราะห์การถดถอย

- ตัวแปรอิสระ X และตัวแปรตาม Y ต้องมีความสัมพันธ์แบบเส้นตรง
- ตัวแปรตามต้องเป็นตัวแปรสุ่มแบบต่อเนื่อง ในขณะที่ตัวแปรอิสระเป็นเซตของค่าต่าง ๆ
- ค่าที่สังเกตได้แต่ละค่าต้องไม่มีความสัมพันธ์กัน
- การแจกแจงความน่าจะเป็นของตัวแปร Y แต่ละค่าของ X ต้องเป็นการแจกแจงแบบปกติ

2.4.2 การวิเคราะห์สหสัมพันธ์ (Correlation Analysis)

ในการวิเคราะห์สหสัมพันธ์เป็นการวัดระดับความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์มากน้อยเพียงใด

ข้อตกลงเบื้องต้นในการวิเคราะห์สหสัมพันธ์

- ตัวแปรอิสระ X และตัวแปรตาม Y ต้องมีความสัมพันธ์แบบเส้นตรง
- ตัวแปรอิสระและตัวแปรตามต้องเป็นตัวแปรสุ่มนชนิดต่อเนื่อง
- ตัวแปรอิสระและตัวแปรตามต้องมีการแจกแจงแบบปกติ
- ตัวแปรแต่ละตัวต้องมีความแปรปรวนเท่ากัน
- ค่าของตัวแปรอิสระและค่าของตัวแปรตามไม่ขึ้นต่อกัน

ในการหาค่าสัมประสิทธิ์สหสัมพันธ์ด้วยวิธีของเพียร์สัน สามารถคำนวณได้จากสูตร

$$r_{xy}=\frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{[N\sum X^2-(\sum X)^2][N\sum Y^2-(\sum Y)^2]}}$$

เมื่อ r_{xy} เป็นค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

$\sum X$ เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)

$\sum Y$ เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)

$\sum XY$ เป็นผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2

$\sum X^2$ เป็นผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1

$\sum Y^2$ เป็นผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2

N เป็นขนาดของกลุ่มตัวอย่าง

(ภัทรศยา ตันติวัฒนกูลและ อรรถกร เก่งพล, 2556)

โดยวิธีของเพียร์สัน มีข้อจำกัด ดังนี้

- ตัวแปรหรือข้อมูลทั้ง 2 ชุด อยู่ในมาตราอันตรภาค หรือมาตราอัตราส่วน
- ข้อมูลทั้ง 2 ชุด มีการแจกแจงแบบปกติ และมีความสัมพันธ์เชิงเส้นตรง
- ข้อมูลในแต่ละชุดจะต้องมีความเป็นอิสระต่อกัน

ตารางที่ 2.4 เกณฑ์การแปลผลความหมายค่าสัมประสิทธิ์ (แสงเดือน วนิชดำรงศักดิ์, 2555) มีดังนี้

ค่าระดับความสัมพันธ์	ระดับความสัมพันธ์
0.81 – 1.00	สูงมาก
0.61 – 0.80	ค่อนข้างสูง
0.41 – 0.60	ปานกลาง
0.21 – 0.40	ค่อนข้างต่ำ
0.01 – 0.20	ต่ำมาก

2.5 การวิเคราะห์หลายตัวแปร

เทคนิคการหาความสัมพันธ์ระหว่าง 2 ตัวแปรขึ้นไปมีหลากหลายวิธี ในการเลือกเทคนิคในการทดสอบควรทราบว่าข้อมูลของเราอยู่ในประเภทใดและสเกลใด เพื่อเลือกใช้เทคนิคในการทดสอบได้อย่างถูกต้อง (กัลยา วานิชย์บัญชา, 2546)

การวิเคราะห์หลายตัวแปรที่จะกล่าวถึง ดังต่อไปนี้

2.5.1 การแบ่งกลุ่มหรือการจำแนกกลุ่มตัวแปร

เทคนิคการวิเคราะห์ปัจจัย (Factor Analysis) เป็นเทคนิคทั่วไปที่ใช้ในการแบ่งกลุ่มหรือการจำแนกกลุ่มตัวแปร โดยเทคนิคการวิเคราะห์ปัจจัยเป็นเทคนิคที่แบ่งตัวแปรออกเป็นกลุ่ม ๆ หรือรวมตัวแปรที่มีความสัมพันธ์กันไว้ในกลุ่มเดียวกัน หรือในปัจจัยเดียวกัน โดยตัวแปรที่อยู่ในปัจจัย (Factor) เดียวกันจะมีความสัมพันธ์กันมาก หากความสัมพันธ์อยู่ในทิศทางเดียวกันจะมีค่าสัมประสิทธิ์สหสัมพันธ์เป็นบวก แต่ถ้าความสัมพันธ์อยู่ในทิศทางตรงกันข้าม จะมีค่าสัมประสิทธิ์

สหสัมพันธ์เป็นลบและ ถ้าตัวแปรที่อยู่ต่างปัจจัยกันจะไม่มีความสัมพันธ์กันหรือ มีความสัมพันธ์กันน้อย ในเทคนิคการวิเคราะห์ปัจจัยจะใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation) วัดความสัมพันธ์ระหว่างตัวแปร ตัวแปรที่จะใช้เทคนิคการวิเคราะห์ปัจจัยได้ควรเป็นตัวแปรเชิงปริมาณ (Interval หรือ Ratio Scale)

เทคนิคการวิเคราะห์ปัจจัยเป็นเทคนิคในการลดจำนวนตัวแปร จากจำนวนตัวแปรมาก ๆ ให้เหลือเพียงไม่กี่ปัจจัย (Factor) โดยจะถือว่าปัจจัยเป็นที่รวมรายละเอียดของตัวแปรอยู่ในปัจจัยนั้น จากนั้นจึงใช้เทคนิคการวิเคราะห์ทางสถิติอื่น ๆ มาวิเคราะห์ปัจจัย โดย 1 ปัจจัย คือ 1 ตัวแปร

2.5.2 การจำแนกกลุ่มข้อมูล

การจำแนกกลุ่มข้อมูลหรือการแบ่งกลุ่มข้อมูลหมายถึง การแบ่งกลุ่มของ คน สัตว์ สิ่งของ เป็นต้น เป็นกลุ่มย่อย ๆ โดยคน สัตว์ สิ่งของ ที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายกัน เช่น รูปร่างลักษณะคล้ายกัน พฤติกรรมคล้ายกัน ส่วนพวกที่อยู่ต่างกลุ่มจะมีความต่างกัน เช่น รูปร่างลักษณะต่างกัน พฤติกรรมต่างกัน โดยเทคนิคที่ใช้ในการจำแนก คน สัตว์ สิ่งของ มี 2 เทคนิค คือ

- Cluster Analysis

เป็นเทคนิคในการแบ่งกลุ่ม Case โดย Case หมายถึง คน สัตว์ สิ่งของ เป็นต้น เป็นกลุ่มย่อย โดย Case ที่มีความคล้ายคลึงกันจะอยู่ในกลุ่มเดียวกัน โดยใช้ตัวแปรที่คาดว่าจะทำให้ Case ต่างกัน ซึ่งเทคนิค Cluster Analysis ไม่จำเป็นต้องทราบจำนวนกลุ่มย่อยว่าควรแบ่งเป็นกี่กลุ่ม และ ไม่ต้องทราบว่า Case ใดอยู่กลุ่มไหนมาก่อน เทคนิคนี้จะจัดกลุ่มโดยศึกษาจากตัวแปรที่นำมาใช้ในการแบ่งกลุ่ม

- Discriminant Analysis

การวิเคราะห์จำแนกกลุ่มหรือ Discriminant Analysis เป็นเทคนิคในการแบ่งกลุ่มข้อมูลหรือ Case หรือหน่วยตัวอย่างเป็นกลุ่มย่อย ๆ หลาย ๆ กลุ่มโดยใช้หลักเกณฑ์การวิเคราะห์ความถดถอยเชิงพหุที่ความสัมพันธ์อยู่ในรูปเชิงเส้น ซึ่งตัวแปรตามเป็นตัวแปรเชิงกลุ่ม ส่วนตัวแปรอิสระเป็นตัวแปรเชิงปริมาณ โดยคนหรือหน่วยที่อยู่ในกลุ่มเดียวกัน จะมีความคล้ายคลึงในตัวแปรที่ศึกษา ซึ่งเทคนิคนี้ต้องทราบมาก่อนว่า แต่ละหน่วยอยู่กลุ่มใดและ มีจำนวนกี่กลุ่ม โดยวัตถุประสงค์ของเทคนิคนี้คือการศึกษาว่าตัวแปรหรือปัจจัยใดบ้างที่เป็นปัจจัยสำคัญทำให้คนหรือหน่วยตัวอย่างอยู่ต่างกลุ่มกัน แล้วนำตัวแปรเหล่านี้มาหาความสัมพันธ์เชิงเส้น จากนั้นนำสมการเชิงเส้นดังกล่าวมาพยากรณ์ หรือประมาณว่าคนหรือหน่วยใหม่ควรอยู่ในกลุ่มใด

2.5.3 การวิเคราะห์ความแปรปรวน

การวิเคราะห์ความแปรปรวนแบบจำแนกทางเดียว (1-Way ANOVA) และการจำแนกความแปรปรวนแบบสองทาง (2-Way ANOVA) ใช้ในกรณีที่ต้องการเปรียบเทียบค่าเฉลี่ยของประชากรตั้งแต่ 3 ประชากรขึ้นไป หรือกล่าวได้ว่า เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามซึ่งเป็นตัวแปรเชิงปริมาณเพียง 1 ตัวและ ตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพที่มีตั้งแต่ 1 ตัวขึ้นไป โดยจะกล่าวถึงการขยายการใช้ ANOVA ดังนี้

- MANOVA (Multivariate Analysis of Variance)

เป็นเทคนิคการวิเคราะห์ความแปรปรวน เมื่อตัวแปรตามเป็นตัวแปรเชิงปริมาณตั้งแต่ 2 ตัวขึ้นไป ส่วนตัวแปรต้นหรือตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพตั้งแต่ 1 ตัวขึ้นไป

- ANCOVA (Analysis of Covariance)

เป็นเทคนิคการวิเคราะห์ความแปรปรวน เมื่อตัวแปรตามเป็นตัวแปรเชิงปริมาณ 1 ตัว และตัวแปรต้นหรือตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ และต้องการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ แต่มีการจำกัด หรือควบคุมตัวแปรเชิงปริมาณที่มีความสัมพันธ์กับตัวแปรตามแฝงอยู่

2.5.4 การวิเคราะห์ความถดถอย

การวิเคราะห์ความถดถอย (Regression Analysis) เป็นเทคนิคในการวิเคราะห์ความสัมพันธ์ของตัวแปรตั้งแต่ 2 ตัวขึ้นไป ซึ่งการวิเคราะห์การถดถอยมีหลายเทคนิคซึ่งขึ้นกับประเภทตัวแปรของข้อมูล

- การวิเคราะห์ความถดถอยโลจิสติกส์ (Logistic Regression Analysis)

การวิเคราะห์ความถดถอยโลจิสติกส์แบบ 2 กลุ่ม (Binary Logistic Regression Analysis) เป็นการวิเคราะห์ความถดถอยที่ไม่ได้อยู่ในรูปเชิงเส้น โดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพทีค่าเพียง 2 ค่า (Dichotomy or Binary Variable) ส่วนตัวแปรอิสระอาจจะเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพหรืออาจมีทั้งตัวแปรเชิงปริมาณและ ตัวแปรเชิงคุณภาพก็ได้ การวิเคราะห์ความถดถอยโลจิสติกส์ไม่มีเงื่อนไขเกี่ยวกับการแจกแจงของตัวแปรอิสระและ เงื่อนไขเกี่ยวกับเมทริกซ์ค่าแปรปรวนและ แปรปรวนร่วมของแต่ละกลุ่ม การวิเคราะห์ความถดถอยโลจิสติกส์แบบ 2 กลุ่ม ในกรณีนี้ตัวแปรตาม Y มีค่าได้เพียง 2 ค่า คือ 0 กับ 1 ดังนั้นตัวแปรตามจะมีการแจกแจงแบบเบอร์นูลี (Bernoulli distribution)

$$P\{Y=y\}=p^{y}(1-p)^{1-y};y=0,1 \qquad (2.1)$$

สำหรับตัวอย่างหน่วยที่ i จะได้ว่า

$$P\{Y_i=y_i\}=p^{y_i}(1-p)^{1-y_i} \; ; y_i = 0,1 \tag{2.2}$$

จากสมการที่ (2.2) เมื่อ $y_i=0$ จะได้

$$P\{Y_i=0\}=p^0(1-p)^{1-0}=1-p$$

เมื่อ $y_i=1$ จะทำให้สมการที่ (2.2) กลายเป็น

$$P\{Y_i=1\}=p^1(1-p)^{1-1}=p$$

$$\begin{aligned} E\{Y_i\}&=\sum Y_i\, P\{Y_i=y_i\} \\ &=0\cdot P\{Y_i=0\}+1\cdot P\{Y_i=1\} \\ \therefore E\{Y\}&=0\cdot(1-p)+1\cdot(p)=p \end{aligned} \tag{2.3}$$

ซึ่งทำให้ $0\leq E\{Y\}\leq 1$

เนื่องจาก Y มีค่าได้เพียง 2 ค่า คือ 0 และ 1 จึงทำให้ความสัมพันธ์ระหว่าง X และY ไม่ได้อยู่ในรูปเชิงเส้น แต่อยู่ในรูป

$$E\{Y\}=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \tag{2.4}$$

โดยเรียกสมการที่ 2.4 ว่า Logistic Response Function โดยที่ $0\leq E\{Y\}\leq 1$
จากสมการที่ 2.3 และ 2.4 จะได้

$$E\{Y\}=p=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{2.5}$$

โดยที่ $p=P\{Y=1\}=P\{$เหตุการณ์ที่สนใจ$\}=E\{Y\}$

$\therefore P\{$ไม่เกิดเหตุการณ์ที่สนใจ$\}=P\{Y=0\}=1-p$

ดังนั้นจากสมการที่ (2.5) จะได้ว่า

$$P\{\text{เกิดเหตุการณ์}\}=P\{Y=1\}=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

หรือ $$P\{Y=1\}=p=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \qquad (2.6)$$

<u>หมายเหตุ</u> การวิเคราะห์ความถดถอยโลจิสติกส์จะต้องใช้ขนาดตัวอย่าง n มากกว่าการวิเคราะห์ความถดถอยแบบปกติ โดยทั่วไป $n \geq 30p$ โดยที่ p เป็นจำนวนตัวแปรอิสระ

วัตถุประสงค์ของการวิเคราะห์ความถดถอยโลจิสติกส์

- เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ พร้อมทั้งศึกษาระดับความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม หรือศึกษาว่าตัวแปรอิสระใดบ้างที่มีอิทธิพลหรือมีผลกระทบต่อตัวแปรตาม
- เพื่อพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจจากสมการที่เหมาะสม โดยการเลือกตัวแปรอิสระที่เหมาะสมเพื่อทำให้เปอร์เซ็นต์ความถูกต้องในการพยากรณ์มีค่าสูงสุด

เงื่อนไขของการวิเคราะห์ความถดถอยโลจิสติกส์ มีดังนี้

- ตัวแปรอิสระ X's อาจเป็นข้อมูลชนิด Dichotomus (มีค่าได้ 2 ค่า) หรือเป็นมาตรอันตรภาค (Interval Scale) และมาตรอัตราส่วน (Ratio Scale) ก็ได้
- ค่าคาดหวังของค่าคลาดเคลื่อนเป็นศูนย์ หรือ E(e) = 0
- e_i และ e_j เป็นอิสระกัน
- e_i และ X_i เป็นอิสระกัน
- ตัวแปรอิสระไม่มีความสัมพันธ์กัน หรือไม่เกิดปัญหา Multicollinearity

2.5.5 การวิเคราะห์ความถดถอยไม่เชิงเส้น (Nonlinear Regression Analysis)

Nonlinear Regression เป็นการวิเคราะห์ความถดถอยเมื่อความสัมพันธ์ของตัวแปรไม่ได้อยู่ในรูปเชิงเส้น

2.6 สารสนเทศร่วม (Mutual Information)

สารสนเทศร่วมได้รับการพัฒนาบนหลักการของทฤษฎีสารสนเทศ (Information Theory) และแนวคิดเรื่องเอนโทรปีที่เสนอโดย Shannon (1948) สมการสารสนเทศร่วมคำนวณได้จากสูตรด้านล่าง

$$MI = \frac{1}{N}\sum_{t=1}^{N} \ln\left[\frac{P_{x,y}(x_i,y_i)}{P_x(x_i),P_y(y_i)}\right]$$

เมื่อ N เป็นขนาดของกลุ่มตัวอย่าง

x_i , y_i เป็นคู่ตัวอย่าง bivariate

$P_{x,y}(x_i,y_i)$ เป็น Join Probability Density ที่จุดตัวอย่าง

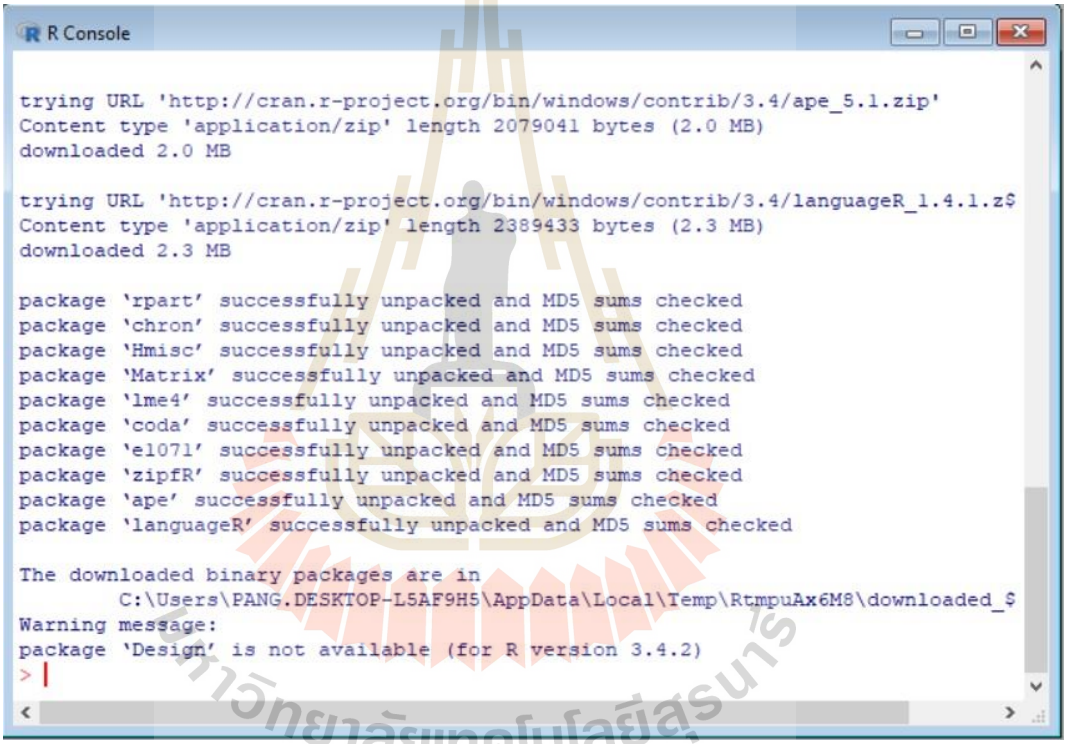
$P_x(x_i) , P_y(y_i)$ เป็น Univariate Marginal Probability Density ที่จุดตัวอย่าง

2.7 โปรแกรม R

โปรแกรม R เป็น Open Source Software ที่ให้ใช้ฟรี สามารถดัดแปลงได้ตามต้องการ ผู้เริ่มต้นเขียนโปรแกรม R คือ Robert Gentleman และ Ross Ihaka (วราฤทธิ์ พานิชกิจโกศลกุล, 2550) โปรแกรม R ได้ถูกเผยแพร่แบบ General Public License ในปี 1995 เป็นโปรแกมที่ใช้ได้ฟรีบนเครื่องคอมพิวเตอร์แบบต่างๆ ไม่ว่าจะใช้บน Windows, Mac OS, หรือ Linux ภาษา R พัฒนามาจากภาษา S ซึ่งพัฒนาขึ้นมาเพื่อใช้ในงานสถิติ (วิโรจน์ อรุณมานะกุล, 2559) โปรแกรม R เป็นทางเลือกหนึ่งที่น่าสนใจ เนื่องจากโปรแกรม R เป็นโอเพนซอร์ส ที่มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูล และมีแพ็กเกจจำนวนมากให้เลือกใช้ได้อย่างสะดวก เป็นโปรแกรมที่มีลิขสิทธิ์ แต่ไม่เสียค่าใช้จ่ายในการใช้งาน สามารถดาวน์โหลดมาใช้ได้ทุกที่ทั่วโลก (กาญจน์ คุ้มทรัพย์, 2559) โดยในโปรแกรม R จะมีแพ็กเกจที่หลากหลายให้ได้เลือกใช้โดยที่ไม่ต้องเขียนโค้ดเอง สามารถดาวน์โหลดแพ็กเกจมาใช้ได้ฟรี เราสามารถติดตั้งแพ็กเกจ เพิ่มเติมในโปรแกรม R ได้ โดยที่แพ็กเกจเป็นชุดคำสั่งหรือฟังก์ชั่นต่างๆ ที่มีผู้เขียนเพิ่มเติมและ ต้องการแบ่งปันให้ผู้อื่นได้ใช้ด้วย สามารถเข้าไปดูที่ https://cran.r-project.org/web/views/ หรือค้นหา แพ็กเกจด้วยคำค้นที่ต้องการใน http://rseek.org/

ในหนังสือ Analyzing Linguistic Data (Baayan 2008) ผู้เขียนได้สร้างแพ็กเกจที่ชื่อ languageR สำหรับใช้ประกอบการอธิบายในหนังสือ เราสามารถติดตั้งแพ็กเกจของหนังสือเล่มนี้โดยพิมพ์คำสั่ง install.packages ตามตัวอย่าง (วิโรจน์ อรุณมานะกุล, 2560) install.packages(c("rpart", "chron", "Hmisc", "Design","Matrix", "lme4", "coda", "e1071", "zipfR", "ape", "languageR"), repos = "http://cran.r-project.org")

โปรแกรมจะเชื่อมต่ออินเทอร์เน็ตไปที่ cran.r-project.org เพื่อ ดาวน์โหลดแพ็กเกจต่างๆที่ระบุ เมื่อเสร็จสิ้นจะเห็นรายงานผล พร้อมทั้งบอกว่าแพ็กเกจนั้นดาวน์โหลดมาไว้ที่ folder ไหน ส่วนตัว แพ็กเกจต่าง ๆ ที่ติดตั้งนั้นจะอยู่ที่ folder c:\Program Files\R\R-2.11.1\library



รูปที่ 2.1 แสดงหน้าต่างหลังจากผู้ใช้ดาวน์โหลดแพ็กเกจเสร็จสิ้น

2.8 โปรแกรม SPSS®

โปรแกรม SPSS® (Statistical Package for the Social Sciences) หรือเรียกอีกชื่อหนึ่งว่า โปรแกรม IBM SPSS Statistics เป็นโปรแกรมสำเร็จรูปสำหรับการวิเคราะห์ข้อมูลทางสถิติ พัฒนาโดยบริษัท SPSS Inc. สหรัฐอเมริกา ในอดีตใช้กับเครื่องคอมพิวเตอร์ชนิดมินิหรือเมนเฟรมคอมพิวเตอร์ ต่อมาได้พัฒนาโปรแกรม SPSS® เป็น SPSS/PC+ ซึ่งใช้กับไมโครคอมพิวเตอร์หรือ คอมพิวเตอร์ส่วนบุคคล โปรแกรม SPSS® เป็นโปรแกรมที่มีประสิทธิภาพสูง ผู้ใช้สามารถวิเคราะห์ข้อมูลโดยใช้สถิติประเภทต่าง ๆ สามารถแสดงผลการวิเคราะห์ได้ทั้งตารางหรือ แผนภูมิชนิดต่าง ๆ การใช้งานของโปรแกรมไม่ซับซ้อน ประมวลผลได้ถูกต้องแม่นยำและ รวดเร็ว

2.9 ปริทัศน์วรรณกรรมที่เกี่ยวข้อง

จากการศึกษาทฤษฎีและทบทวนวรรณกรรมสรุปได้ว่า การหาค่าความสัมพันธ์ของข้อมูลมีหลากหลายวิธีและ เงื่อนไขในการใช้ที่แตกต่างกันไป ซึ่งแต่ละวิธีอาจเหมาะสมกับแต่ละประเภทของข้อมูลนั้น ๆ แต่มีวัตถุประสงค์เดียวกันคือ เป็นการหาความสัมพันธ์ระหว่างข้อมูล ดังนั้นก่อนทำการเลือกวิธีการ ผู้วิจัยต้องทราบประเภทของข้อมูลในการวิจัยเสียก่อนและ ดูเงื่อนไขในการหาความสัมพันธ์ของแต่ละวิธี แล้วจึงสามารถเลือกวิธีการในการหาค่าความสัมพันธ์

จากการศึกษางานวิจัยที่เกี่ยวข้องด้านการวิเคราะห์ความสัมพันธ์ของข้อมูลที่มีผลต่อคุณภาพของฮาร์ดดิสก์ พบวิธีวิจัยที่หลากหลายคือ อโณทัย ศิลเทพาเวทย์ (2554) ได้ทำการปรับปรุงคุณภาพผลิตภัณฑ์ในการผลิตฮาร์ดดิสก์โดยวิธีแผนภูมิต้นไม้ (Decision Three) โดยค้นคว้าวิธีการเชิงระบบ (System Asthmatic Approach) เพื่อหาพารามิเตอร์ที่เหมาะสมในกระบวนการผลิตฮาร์ดดิสก์และ เลือกอัลกอลิทึมที่เหมาะสมในการปรับปรุงพารามิเตอร์ เพราะไม่สามารถทดสอบผลลัพธ์ได้ในสภาพแวดล้อมจริง เขาจึงเสนอวิธีการที่น่าเชื่อถือในการทำนายเพื่อปรับปรุงพารามิเตอร์ให้ดีมากขึ้น จากการศึกษาทำให้สภาพสินค้ามีของเสียน้อยลง 12% จึงเลือกวิธีการเชิงระบบมาปรับปรุงคุณภาพในการผลิตฮาร์ดดิสก์ ต่อมา ธนดล สุชาติพงศ์ (2557) ทำเหมืองข้อมูลโดยเลือกแอตทริบิวท์และ ปรับแต่งข้อมูลที่ได้มาจากกระบวนการทดสอบคุณภาพ จากนั้นใช้อัลกอริทึม C5.0, Neural Network, C&R Tree, SVM และ CHAID ในการเรียนรู้ประเภทของเสียของฮาร์ดดิสก์ ซึ่งเขาเปรียบเทียบผลลัพธ์ในแต่ละวิธีที่สามารถคัดแยกการเสียของฮาร์ดดิสก์ได้แม่นยำมากที่สุดซึ่งพบว่า แบบจำลองอัลกอริทึม C5.0 ให้ผลลัพธ์ได้แม่นยำมากที่สุด โดยมีค่าความถูกต้อง 99.79% และเนื่องจากข้อมูลของกระบวนการผลิตฮาร์ดดิสก์มีขนาดใหญ่ จามรี ชูบัวทองและ สมศรี บัณฑิตวิไล (2560) ก็ได้ใช้วิธีเหมืองข้อมูลเช่นเดียวกันแต่จามรี ชูบัวทอง และสมศรี บัณฑิตวิไล จะใช้

เทคนิคการวิเคราะห์การถดถอยโลจิสติกส์ เพื่อศึกษาตัวแปรอิสระที่ส่งผลกระทบต่อคุณภาพของฮาร์ดดิสก์ ส่วนในด้านการใช้ Mutual Information (MI) ได้ค้นพบงานวิจัยของ Luis M. de Campos (2006), Nara Samattapapong (2559) และ Haodi Jiang et al. (2017) เข้ามาใช้ในการหาความสัมพันธ์ของข้อมูล โดยงานวิจัยของ Luis M. de Campos (2006) ได้ทำฟังก์ชั่นการให้คะแนนสำหรับเรียนรู้ Bayesian Networks บนพื้นฐานของ Mutual Information และ การทดสอบเงื่อนไขความเป็นอิสระของข้อมูล โดยเขาพัฒนาฟังก์ชั่นการให้คะแนนใช้ Mutual Information เพื่อวัดระดับความสัมพันธ์ระหว่างตัวแปรมาใช้ในการค้นหาโครงสร้างเครือข่ายที่ดีที่สุด ซึ่งผลการทดลองของพวกเขาให้ประสิทธิภาพมากกว่าวิธีการอนุมานเครือข่ายอนุกรมเวลาแบบอื่น ๆ ยังพบงานวิจัยของ Nara Samattapapong (2559) ได้ใช้ Mutual Information (MI) เป็นส่วนหนึ่งในการวิเคราะห์ข้อมูลก่อนการพยากรณ์คุณภาพของฮาร์ดดิสก์ โดยเขาได้อธิบายว่า Mutual Information สามารถเลือกตัวแปรที่มีความสัมพันธ์เชิงเส้นหรือไม่เชิงเส้น การกำหนดความสัมพันธ์ของข้อมูลก่อนการเลือกใช้จึงไม่จำเป็น วิธีนี้สามารถตรวจสอบชุดข้อมูลทั้งหมดได้อย่างรวดเร็วโดยไม่ต้องมีการประมวลผลข้อมูลก่อนและ Haodi Jiang et al. (2017) ได้ศึกษาเกี่ยวกับวิศวกรรมย้อนกลับ GRNs และ RNMs ของเซลล์ โดยใช้ Dynamic Bayesian Network และฟังก์ชั่นการให้คะแนน Mutual Information เพื่ออนุมาน GRNs และ RNMs ของเซลล์ได้โดยอัตโนมัติจาก time series ของชุดข้อมูล พวกเขาได้นำวิธีฟังก์ชั่นการให้คะแนนด้วยวิธี Mutual Information ของ Luis M. de Campos (2006) เข้ามาใช้ร่วมด้วย

ด้านการศึกษางานวิจัยที่ใช้โปรแกรมในการวิเคราะห์ข้อมูลทางสถิติพบว่า นิรมล พันสีมา และ อนัตต์ เจ่าสกุล, 2557 เปรียบเทียบการทำงานของโปรแกรม R และโปรแกรม SPSS® ในการวิเคราะห์ปัจจัยจากฐานข้อมูลเงินยืมทดรองจ่ายของมหาวิทยาลัยขอนแก่น โดยสร้างโมเดลจำแนกการยืมคืนเงินยืมทดรองจ่ายโดยใช้โปรแกรม R และ SPSS® ซึ่งวิเคราะห์เปรียบเทียบการทำงานและ ประสิทธิภาพของโปรแกรมจากผลการวิจัยพบว่า โปรแกรม R มีความยืดหยุ่นของรูปแบบไฟล์ข้อมูลที่นำเข้ามากกว่าและ มีการจัดการข้อมูลที่รวดเร็ว ใช้งานง่ายโดยโปรแกรม SPSS® มีข้อจำกัดในการทำงานและ วิเคราะห์ข้อมูลค่อนข้างมากเมื่อเทียบกับโปรแกรม R ที่สามารถวิเคราะห์ข้อมูลสถิติขั้นสูงได้และ ยังสามารถวิเคราะห์ข้อมูลได้หลากหลายรูปแบบ

จากการศึกษางานวิจัยยังไม่พบงานวิจัยที่ใช้โปรแกรม R ในการศึกษาความสัมพันธ์ของข้อมูลที่มีผลต่อคุณภาพของฮาร์ดดิสก์ ซึ่งผู้วิจัยจึงสนใจโปรแกรม R มาเป็นเครื่องมือในการวิจัย เนื่องจากโปรแกรม R เป็นโปรแกรมทางสถิติที่ให้ใช้ฟรีและ มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูล ซึ่งสามารถช่วยลดค่าใช้จ่ายของบริษัทกรณีศึกษาในเรื่องของลิขสิทธิ์ซอฟต์แวร์กล่าวดังบทที่

บทที่ 3
วิธีดำเนินการวิจัย

ในการศึกษางานวิจัย เรื่อง การวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพงานในกระบวนการทดสอบฮาร์ดดิสก์ เพื่อศึกษาความสัมพันธ์ของตัวแปรอิสระต่างๆ ที่มีผลต่อคุณภาพของฮาร์ดดิสก์ ในบทนี้จะกล่าวถึงรูปแบบการดำเนินการวิจัย ตัวอย่างข้อมูล การวิเคราะห์ข้อมูล สถิติที่ใช้ในการวิจัย และการดำเนินการวิจัย โดยมีรายละเอียดดังต่อไปนี้

- ประชากรและกลุ่มตัวอย่าง
- กรอบแนวคิดและตัวแปรที่ใช้ในการวิจัย
- วิธีการวิเคราะห์ข้อมูล
- สถิติที่ใช้ในการวิจัย
- การดำเนินการวิจัย

3.1 ประชากรและกลุ่มตัวอย่าง

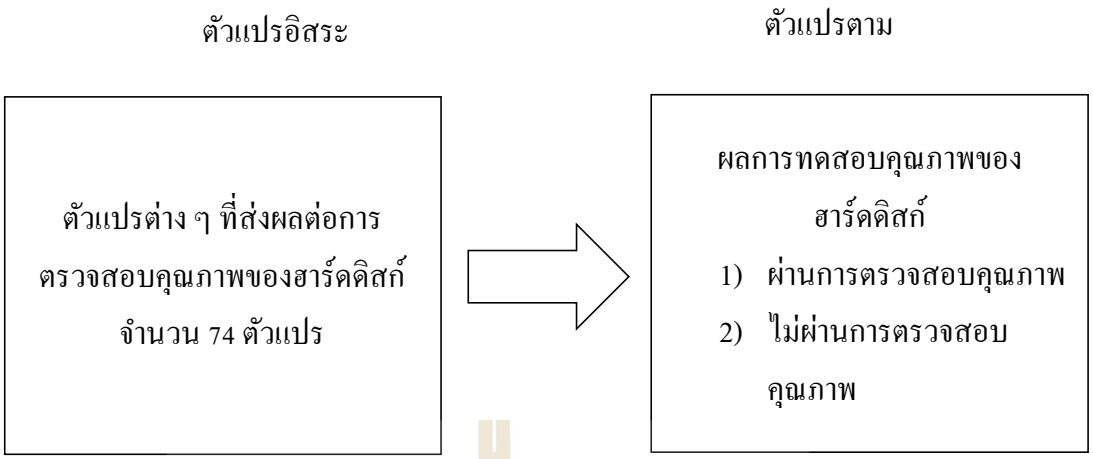
ข้อมูลในการวิจัยครั้งนี้มาจากบริษัทกรณีศึกษา โดยข้อมูลที่ได้มาจากขั้นตอนกระบวนการทดสอบคุณภาพของฮาร์ดดิสก์ก่อนส่งถึงมือลูกค้า ซึ่งข้อมูลที่ได้รับมามีจำนวนทั้งหมด 57,232 ตัว

3.2 กรอบแนวคิดและตัวแปรที่ใช้ในการวิจัย

3.2.1 กรอบแนวคิด (Conceptual Framework)

ในการวิจัยมีขอบเขตการวิจัย ดังนี้

ประชากร คือ ผลการทดสอบคุณภาพของฮาร์ดดิสก์จำนวนทั้งหมด 57,232 ตัว โดยมีกรอบแนวคิดในการวิจัย ดังนี้



รูปที่ 3.1 กรอบแนวคิดงานวิจัย

3.2.2 ตัวแปร (Variables)

ตัวแปรตาม คือ ผลการทดสอบคุณภาพของฮาร์ดดิสก์ (ผ่านการตรวจสอบคุณภาพ/ไม่ผ่านการตรวจสอบคุณภาพ)

ตัวแปรอิสระ คือ ... (ไม่สามารถเปิดเผยความหมายของตัวแปรอิสระแต่ละตัวได้ ซึ่งเป็นความลับของบริษัทกรณีศึกษา)

3.3 วิธีการวิเคราะห์ข้อมูล

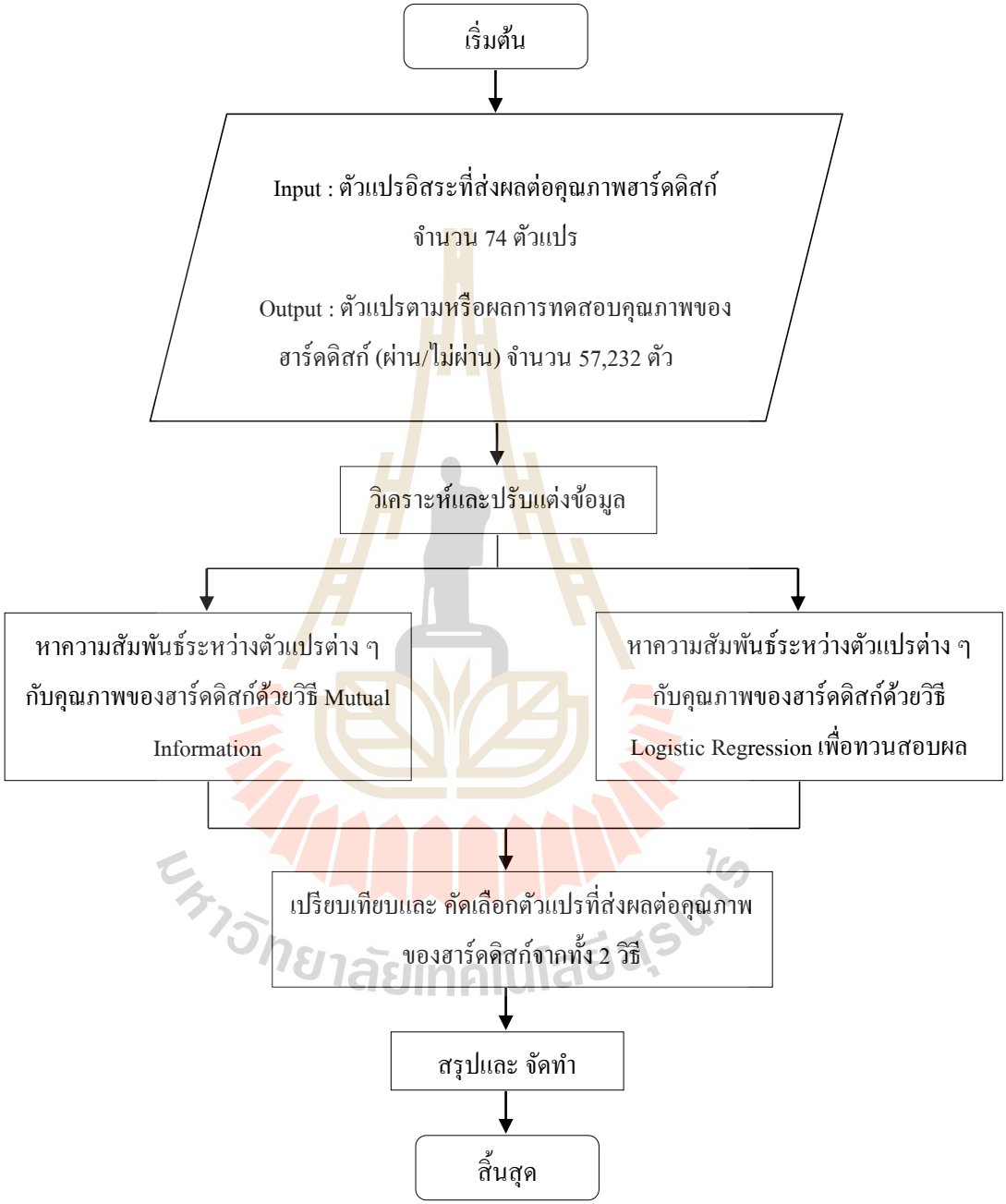
ผู้วิจัยได้นำข้อมูลผลการทดสอบคุณภาพของฮาร์ดดิสก์และตัวแปรต่าง ๆ ที่ส่งผลต่อคุณภาพฮาร์ดดิสก์ที่ได้รับจากบริษัทกรณีศึกษา วิเคราะห์ข้อมูลด้วยโปรแกรม R และทวนสอบผลด้วยโปรแกรม SPSS

3.4 สถิติที่ใช้ในการวิจัย

ในการวิจัยนี้จะศึกษาความสัมพันธ์ระหว่างตัวแปร 2 วิธี ได้แก่

- Mutual Information
- Logistic Regression

งานวิจัยเรื่อง การวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพงานในกระบวนการทดสอบฮาร์ดดิสก์ มีขั้นตอนการดำเนินงานวิจัย แสดงดังรูปที่ 3.2



รูปที่ 3.2 แสดง (Flow Chart) ขั้นตอนการดำเนินงานวิจัย

3.5 การดำเนินการวิจัย

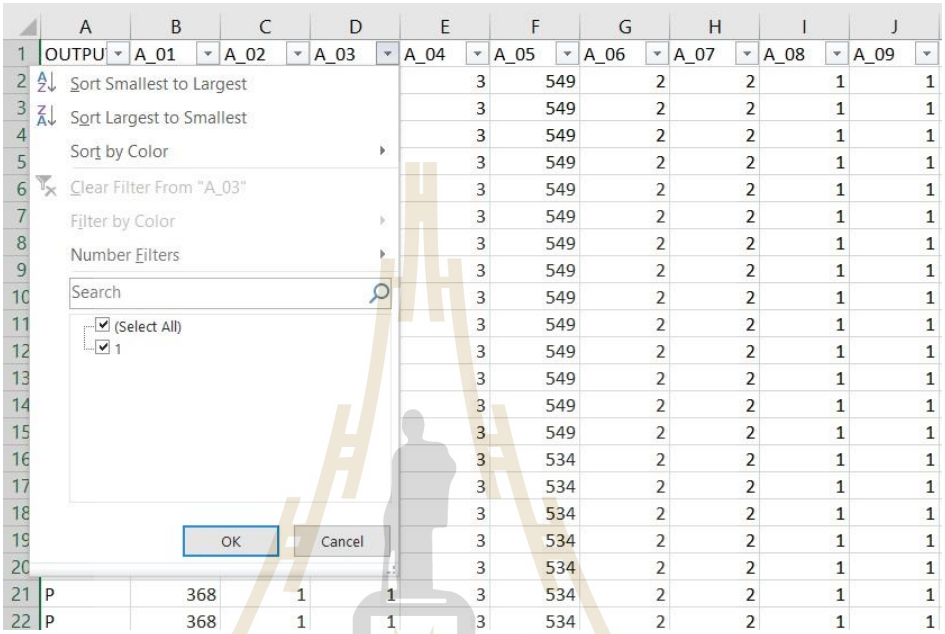
ในส่วนของการดำเนินการวิเคราะห์ความสัมพันธ์ของตัวแปรในกระบวนการผลิตที่มีผลต่อคุณภาพงานในกระบวนการทดสอบฮาร์ดดิสก์ ผู้วิจัยได้ทำการปรับแต่งข้อมูลก่อนการเข้าสู่การวิเคราะห์ข้อมูล จากนั้นจึงดำเนินการวิเคราะห์ความสัมพันธ์โดยโปรแกรม R ด้วยวิธี Mutual Information และทวนสอบผลการวิเคราะห์ความสัมพันธ์โดยโปรแกรม SPSS ด้วยวิธี Logistic Regression โดยแสดงขั้นตอนการดำเนินการวิจัยในหัวข้อที่ 3.5.1, 3.5.2 และ 3.5.3 ตามลำดับ

3.5.1 การปรับแต่งข้อมูลก่อนเข้าสู่การวิเคราะห์ค่าความสัมพันธ์ของข้อมูล

ข้อมูลที่ได้จากบริษัทกรณีศึกษาซึ่งวัดค่าได้จากกระบวนการตรวจสอบคุณภาพของฮาร์ดดิสก์ โดยข้อมูลที่ได้รับมีจำนวนตัวแปรคุณภาพของฮาร์ดดิสก์ทั้งหมด 57,232 ตัว และจำนวนตัวแปรที่เกี่ยวข้องต่อคุณภาพของฮาร์ดดิสก์ 74 ตัว ซึ่งแบ่งเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจำนวน 53 ตัว และตัวแปรเชิงปริมาณจำนวน 21 ตัว ก่อนทำการวิจัยผู้วิจัยได้ทำการลบแถวของคุณภาพฮาร์ดดิสก์ที่ไม่มีผลคุณภาพของฮาร์ดดิสก์ว่าผ่านหรือไม่ผ่านการตรวจสอบคุณภาพและคุณภาพของฮาร์ดดิสก์ที่มีตัวแปรไม่ครบทั้ง 74 ตัวแปรออก ซึ่งจะเหลือข้อมูลจำนวนตัวแปรคุณภาพของฮาร์ดดิสก์ในการคำนวณ 39,605 ตัว และจำนวนตัวแปรที่เกี่ยวข้องต่อคุณภาพของฮาร์ดดิสก์ 74 ตัวดังเดิม

ผู้วิจัยได้ทำการกรองข้อมูล เพื่อดูค่าของแต่ละตัวแปรอิสระว่ามีค่าเป็นอย่างไร เมื่อทำการกรองข้อมูลในโปรแกรม Excel พบว่าตัวแปรอิสระเชิงคุณภาพ (Attribute Variables) ต่อไปนี้มีค่าเพียง 1 ค่า โดยตัวแปรอิสระ A_03, A_08, A_09, A_10, A_14, A_15, A_41, A_44, A_45, A_46, A_61 และ A_71 มีค่าเท่ากับ 1 ในทุก ๆ ของตัวแปรตาม กล่าวคือ ไม่ว่าคุณภาพของฮาร์ดดิสก์จะเป็น 0 หรือ 1 ค่าของตัวแปร A_03, A_08, A_09, A_10, A_14, A_15, A_41, A_44, A_45, A_46, A_61, A_71 จะมีค่าเท่ากับ 1 เสมอและ ยังพบอีกว่าตัวแปรอิสระ A_29, A_30, A_31 และ A_32 มีค่าเท่ากับ 2 เสมอ ไม่ว่าคุณภาพของฮาร์ดดิสก์จะเป็น 1 หรือ 0 โดยที่คุณภาพของฮาร์ดดิสก์ที่มีค่าเท่ากับ 1 คือ ฮาร์ดดิสก์ที่ผ่านขั้นตอนการตรวจสอบคุณภาพและ ฮาร์ดดิสก์ที่มีค่าเท่ากับ 0 คือ ฮาร์ดดิสก์ที่ไม่ผ่านขั้นตอนการตรวจสอบคุณภาพ ซึ่งตัวแปรอิสระเชิงคุณภาพดังกล่าวข้างต้น มีค่าของตัวแปรเพียงค่าเดียว ในการวิเคราะห์หรือการหาค่าสัมประสิทธิ์สหสัมพันธ์ไม่สามารถคำนวณหาคู่ของตัวแปรได้เพราะ ตัวแปรมีความแปรปรวนเป็น 0 จึงต้องทำการตัดตัวแปรอิสระเชิงคุณภาพออกเป็นจำนวน 16 ตัวแปร ก็คือ A_03, A_08, A_09, A_10, A_14, A_15, A_41, A_44, A_45, A_46, A_61, A_71, A_29, A_30, A_31 และ A_32 ดังนั้นจะเหลือตัวแปรอิสระในการคำนวณ 58 ตัวแปร คือ A_01, A_02, A_04, A_05, A_06, A_07, A_11, A_12, A_13, A_16, A_17, A_18, A_19, A_20, A_21, A_22, A_23, A_24, A_25, A_26, A_27, A_28, C_33, C_34, C_35,

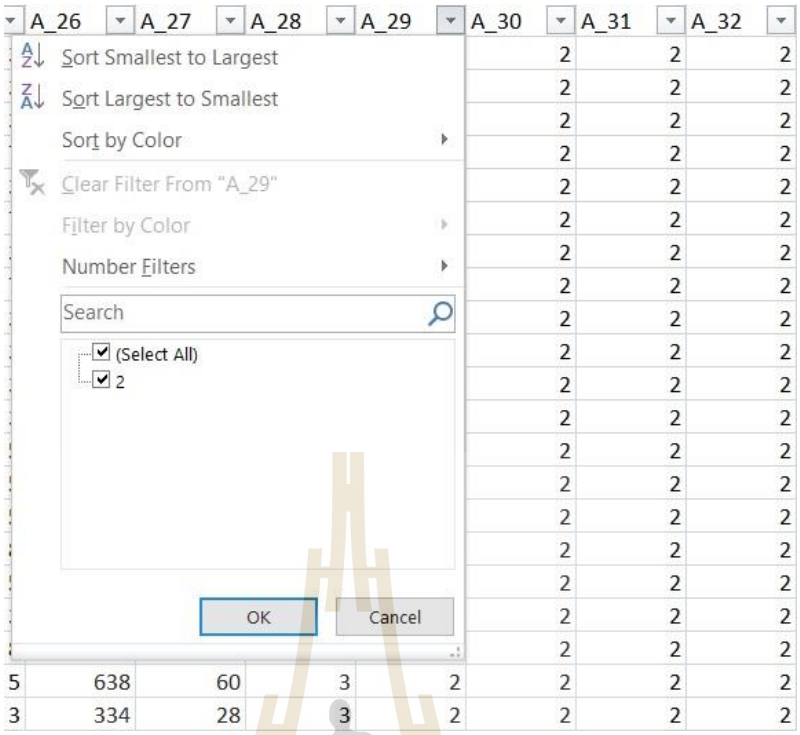
C_36, C_37, C_38, A_39, A_40, A_42, A_43, C_47, C_48, C_49, C_50, C_51, A_52, C_53, A_54, C_55, C_56, C_57, C_58, C_59, C_60, A_62, A_63, A_64, A_65, C_66, C_67, C_68, A_69, A_70, A_72, A_73 และ A_74 โดยจะแสดงตัวอย่างการกรองข้อมูลก่อนการวิเคราะห์ความสัมพันธ์ดังรูปที่ 3.3 และ 3.4



	A	B	C	D	E	F	G	H	I	J
1	OUTPU	A_01	A_02	A_03	A_04	A_05	A_06	A_07	A_08	A_09
2					3	549	2	2	1	1
3					3	549	2	2	1	1
4					3	549	2	2	1	1
5					3	549	2	2	1	1
6					3	549	2	2	1	1
7					3	549	2	2	1	1
8					3	549	2	2	1	1
9					3	549	2	2	1	1
10					3	549	2	2	1	1
11					3	549	2	2	1	1
12					3	549	2	2	1	1
13					3	549	2	2	1	1
14					3	549	2	2	1	1
15					3	549	2	2	1	1
16					3	534	2	2	1	1
17					3	534	2	2	1	1
18					3	534	2	2	1	1
19					3	534	2	2	1	1
20					3	534	2	2	1	1
21	P	368	1	1	3	534	2	2	1	1
22	P	368	1	1	3	534	2	2	1	1

รูปที่ 3.3 แสดงการกรองข้อมูลที่ตัวแปร A_01 เพื่อดูค่าของข้อมูลเท่ากับ 1

จากรูปที่ 3.3 แสดงการกรองข้อมูลที่ตัวแปร A_01 เพื่อดูค่าของข้อมูลซึ่งค่าของตัวแปรมีค่าเท่ากับ 1 เสมอ ไม่ว่าคุณภาพของฮาร์ดดิสก์จะผ่านหรือ ไม่ผ่านกระบวนการทดสอบ



รูปที่ 3.4 แสดงการกรองข้อมูลที่ตัวแปร A_01 เพื่อดูค่าของข้อมูลเท่ากับ 2

จากรูปที่ 3.4 แสดงการกรองข้อมูลที่ตัวแปร A_29 เพื่อดูค่าของข้อมูล ซึ่งค่าของตัวแปรมีค่าเท่ากับ 2 เสมอ ไม่ว่าคุณภาพของฮาร์ดดิสก์จะผ่านหรือ ไม่ผ่านกระบวนการทดสอบ

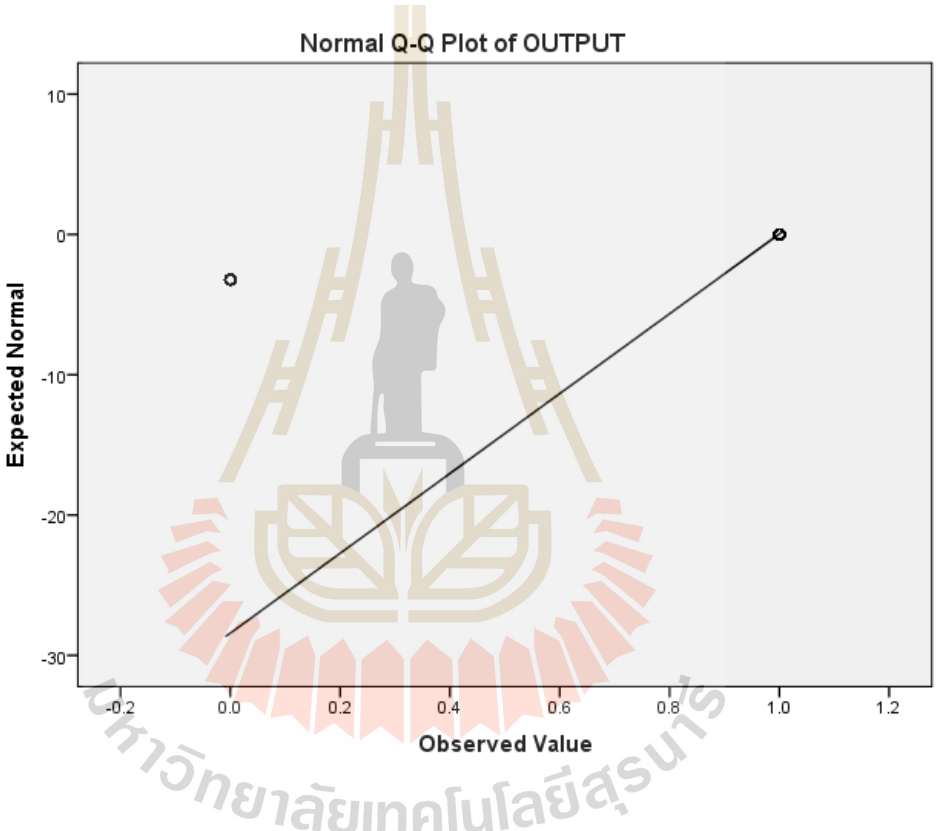
จากนั้นทำการเปลี่ยนค่าของคุณภาพฮาร์ดดิสก์ โดยเปลี่ยนข้อมูลตัวอักษรเป็นข้อมูลตัวเลขคือ ให้ข้อมูลตัวอักษร P เปลี่ยนเป็นข้อมูลตัวเลข 1 หมายถึง ฮาร์ดดิสก์ผ่านกระบวนการตรวจสอบคุณภาพและ ข้อมูลตัวอักษร F เปลี่ยนเป็นข้อมูลตัวเลข 0 หมายถึง ฮาร์ดดิสก์ไม่ผ่านกระบวนการตรวจสอบคุณภาพ เพื่อให้เหมาะสมต่อการคำนวณ ผู้วิจัยได้ทำการทดสอบ Normal Distribution ของตัวแปรตามและ ตัวแปรอิสระเพื่อดูว่าการแจกแจงเป็นแบบปกติหรือไม่ โดยผลการทดสอบการกระจายตัวของตัวแปรอิสระมีทั้งการกระจายตัวแบบปกติและไม่ปกติ ซึ่งการกระจายตัวดังกล่าวไม่สามารถใช้วิธีสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันได้ เนื่องจากวิธีสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันมีข้อจำกัดคือ ข้อมูลทั้ง 2 ชุด ต้องมีการแจกแจงแบบปกติ ซึ่งกล่าวแล้วในบทที่ 2 หัวข้อที่ 2.4.2

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
OUTPUT	.513	39605	.000

a. Lilliefors Significance Correction

รูปที่ 3.5 แสดงค่า Significance ของตัวแปรตาม (OUTPUT)



รูปที่ 3.6 แสดงการกระจายตัวของตัวแปรตาม

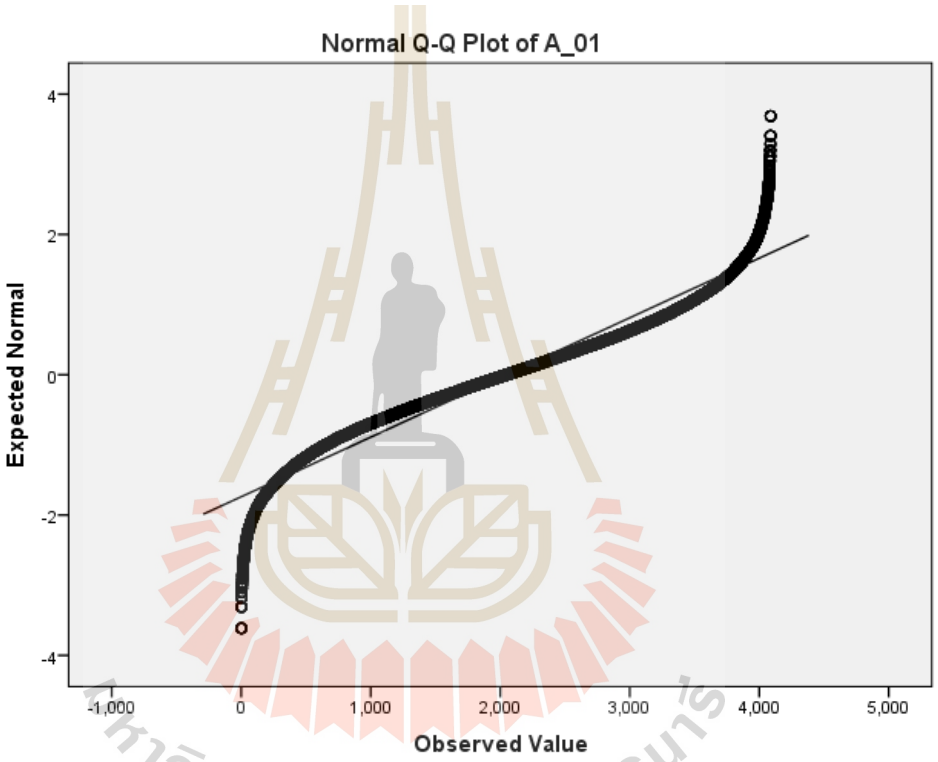
จากรูปที่ 3.5 และ 3.6 พิจารณาการแจกแจงแบบปกติด้วยสถิติ Kolmogorov-Smirnov Test พบว่าค่า Significance เท่ากับ .000 น้อยกว่าค่าแอลฟา (< .05) แปลว่า ปฏิเสธ H0 สรุปข้อมูลนี้มีการแจกแจงแบบไม่ปกติ เนื่องจากตัวแปรตาม Y มีค่าเพียง 2 ค่า คือ 0 กับ 1 ดังนั้นตัวแปรตามจะมีการแจกแจงแบบเบอร์นูลี (Bernoulli Distribution) (กัลยา วานิชย์บัญชา, 2546)

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
A_01	.057	39605	.000

a. Lilliefors Significance Correction

รูปที่ 3.7 แสดงค่า Significance ของตัวแปรอิสระ A_01



รูปที่ 3.8 แสดงการกระจายตัวของตัวแปรอิสระ (A_01)

จากรูปที่ 3.7 และ 3.8 พิจารณาการแจกแจงแบบปกติด้วยสถิติ Kolmogorov-Smirnov Test พบว่าค่า Significance เท่ากับ .000 น้อยกว่าค่าแอลฟา (< .05) แปลว่า ปฏิเสธ H0 สรุปข้อมูลนี้มีการแจกแจงแบบไม่ปกติ

จากการศึกษาทฤษฎีต่าง ๆ ที่เกี่ยวข้อง ผู้วิจัยสามารถสรุปตารางการใช้สถิติวิเคราะห์ความสัมพันธ์ของตัวแปร 2 ตัว ได้ดังตารางที่ 3.1

ตารางที่ 3.1 สรุปการวิเคราะห์ความสัมพันธ์ของตัวแปร 2 ตัว

ตัวแปร	ชนิดของข้อมูล	วิธีการ								
		สัมประสิทธิ์สหสัมพันธ์เพียร์สัน	สัมประสิทธิ์สหสัมพันธ์สเปียร์แมน	gamma	t-Test	Z-Test	F-Test	การทดสอบไคสแควร์	Simple Regression	Logistic Regression
ตัวแปรอิสระ	เชิงคุณภาพ	✗	✓	✓	✓	✓	✓	✓	✗	✓
	เชิงปริมาณ	✓	✗	✗	✗	✗	✗	✗	✓	✓
ตัวแปรตาม	เชิงคุณภาพ	✓	✓	✓	✗	✗	✗	✓	✗	✓
	เชิงปริมาณ	✗	✗	✗	✓	✓	✓	✗	✓	✗

จากตารางที่ 3.1 พิจารณาได้ว่าวิธีโลจิสติกส์เหมาะสมกับการวิเคราะห์ข้อมูลของงานวิจัยนี้มากที่สุดเพราะ งานวิจัยนี้มีตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ในขณะที่ตัวแปรอิสระเป็นทั้งตัวแปรเชิงคุณภาพและเชิงปริมาณ

3.5.2 ขั้นตอนการดำเนินการวิเคราะห์ความสัมพันธ์โดยโปรแกรม R ด้วยวิธี Mutual Information

ในขั้นตอนการวิเคราะห์ความสัมพันธ์โดยโปรแกรม R ด้วยวิธี Mutual Information ผู้วิจัยได้ศึกษาแพ็กเกจที่ช่วยในการคำนวณค่า MI ซึ่งช่วยลดเวลาในการดำเนินงานวิจัยได้เพราะ ไม่ต้องทำการเขียนโปรแกรมในการคำนวณเองให้ยุ่งยาก ได้พบงานวิจัยของ Patrick E. Meyer, 2014 พัฒนา 'infotheo' ซึ่งเป็นแพ็กเกจหนึ่งที่สามารถติดตั้งหลังจากติดตั้งโปรแกรม R ได้ในทันที 'infotheo' ถูกนำมาใช้เป็นแพ็กเกจ R ซึ่งแพ็กเกจ 'infotheo' สามารถใช้งานได้อิสระบน CRAN โดยที่ผู้ใช้งานไม่ต้องเขียนโปรแกรมในการคำนวณเอง ซึ่งเขาได้อธิบายเกี่ยวกับแพ็กเกจนี้ว่า "แพ็กเกจนี้ใช้มาตรการต่าง ๆ ของทฤษฎีสารสนเทศโดยอิงตามการประมาณค่าเอนโทรปีหลายตัว ใช้ตัวแปรสุ่มสองตัวแปรเป็นข้อมูลเข้าและคำนวณข้อมูลร่วมกันตามวิธีการประมาณเอนโทรปี ถ้า Y ไม่ได้จัดมาให้และ X เป็นเมทริกซ์, ฟังก์ชันจะส่งกลับเมทริกซ์ของข้อมูลร่วมกันระหว่างทุกคู่ของตัวแปรในชุดข้อมูล X

ในตัวแพ็กเกจ 'infotheo' มีหลายฟังก์ชันในการคำนวณค่าต่าง ๆ ซึ่งผู้วิจัยได้เลือกใช้ฟังก์ชัน mutinformation ในการคำนวณค่า MI Value ซึ่ง Patrick E. Meyer ได้อธิบายการใช้ mutinformation คือ mutinformation(X, Y, method="emp") โดยที่ X คือ vector/factor แสดงถึงตัวแปรสุ่มหรือ กรอบข้อมูลแสดงถึงเวคเตอร์สุ่ม ซึ่งคอลัมน์นั้นประกอบด้วยตัวแปร/คุณสมบัติ และแถวประกอบด้วยผลลัพธ์/ตัวอย่าง Y คือ ตัวแปรสุ่มหรือเวคเตอร์สุ่มอีกตัวแปรหนึ่ง" โดยขั้นตอนการหาค่า Mutual Information แสดงดังขั้นตอนที่ 1) - 11)

1) เปิดโปรแกรม R (ดังรูปที่ 3.9) (รายละเอียดวิธีการติดตั้งโปรแกรม R ตามภาคผนวก ก)

2) ทำการติดตั้งแพ็กเกจ (Package Infotheo) ก่อนการคำนวณ โดยไปที่ Packages > Install Package(s) > (HTTP mirror) > Thailand > Infotheo (ดังรูปที่ 3.10)

3) เมื่อทำการติดตั้งแพ็กเกจเรียบร้อยแล้ว ทำการโหลดแพ็กเกจ โดยไปที่ Packages > Load Package(s) > Infotheo (ดังรูปที่ 3.11)

4) โปรแกรมจะขึ้นหน้าต่างการใช้งานของแพ็กเกจ Infotheo โดยให้ใส่คำสั่งในช่อง R Console (ดังรูปที่ 3.12)

5) ทำการนำเข้าข้อมูล excel เข้ามาในโปรแกรม R โดยต้องบันทึกไฟล์ excel เป็นสกุล .csv ก่อน (เนื่องจากวิธีนี้นำเข้าได้แต่สกุล .csv) โดยพิมพ์คำสั่ง ดังนี้

```
library(readr)
data <- read_csv("ที่อยู่ของไฟล์/ชื่อไฟล์ excel.csv") (ดังรูปที่ 3.13)
```

ตัวอย่างคำสั่ง

```
library(readr)
data <- read_csv("C:/Users/User/Desktop/senior project/rawdata Rstudio.csv")
```

6) พิมพ์คำสั่งที่ใช้คำนวณค่า MI ดังนี้

mutinformation(data$ชื่อคอลัมน์ตัวแปรอิสระ, data$ชื่อคอลัมน์ตัวแปรตาม, method="emp") (ดังรูปที่ 3.14)

ตัวอย่างคำสั่ง

```
mutinformation(data$A_01, data$OUTPUT, method="emp")
```

7) พิมพ์คำสั่งเดิมแต่เปลี่ยนชื่อคอลัมน์จนครบจำนวนคอลัมน์ของตัวแปรอิสระที่มี กล่าวคือเปลี่ยนคอลัมน์ data$A_01 เป็น data$A_02 ไปเรื่อย ๆ จนครบทุกชื่อตัวแปรอิสระทั้ง 74 ตัว (ดังรูปที่ 3.15)

ตัวอย่างคำสั่ง

```
mutinformation(data$A_02, data$OUTPUT, method="emp")
mutinformation(data$A_03, data$OUTPUT, method="emp")
```

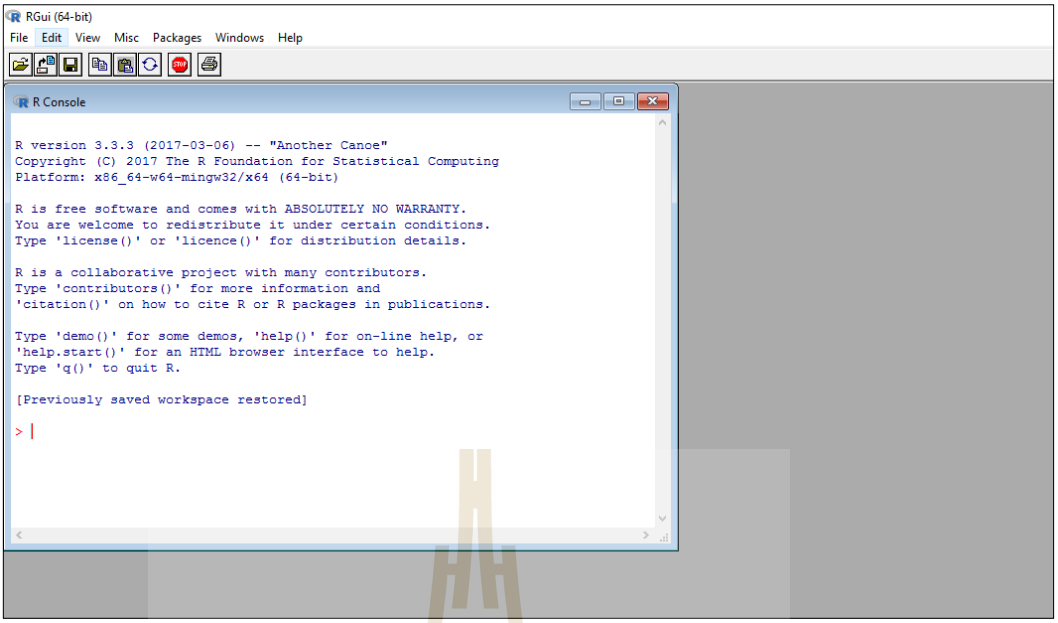
เปลี่ยนคอลัมน์ Rawdata$A_01 เป็น Rawdata$A_02 ไปเรื่อย ๆ จนครบทุกชื่อตัวแปรอิสระทั้ง 74 ตัว

8) เขียนโค้ด VBA เพื่อวนลูปชื่อตัวแปรอิสระทั้ง 74 ตัว (ดังรูปที่ 3.16)

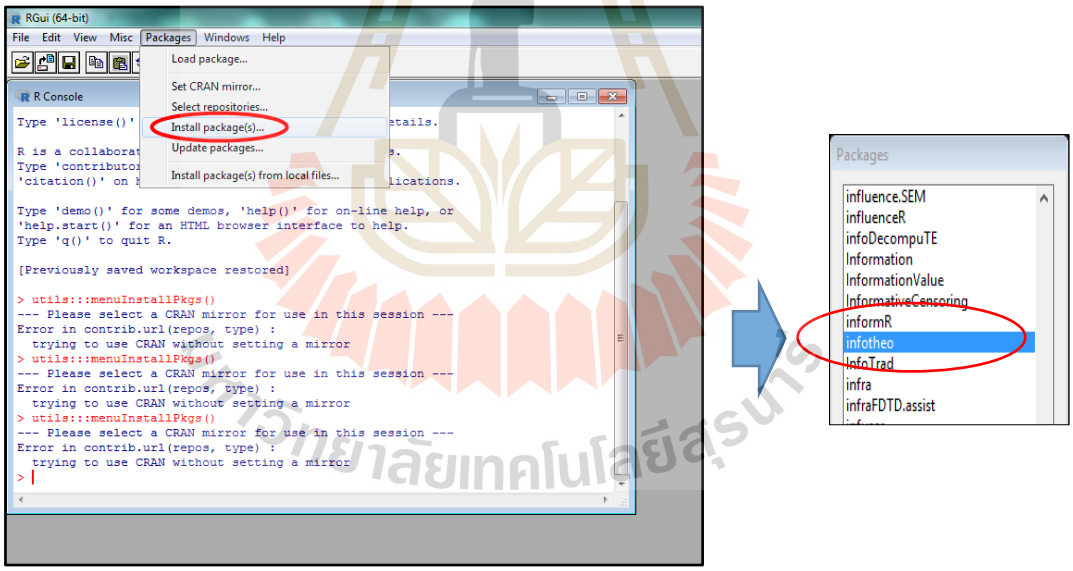
9) ทำการคัดลอกผลลัพธ์ที่ได้จากการรัน VBA (ดังรูปที่ 3.17) ใส่ในการหน้าต่างของการหาค่า MI ในโปรแกรม R (ดังรูปที่ 3.18)

10) ทำการกด Enter จะได้ผลลัพธ์ของ MI Value ทั้งหมด (ดังรูปที่ 3.19)

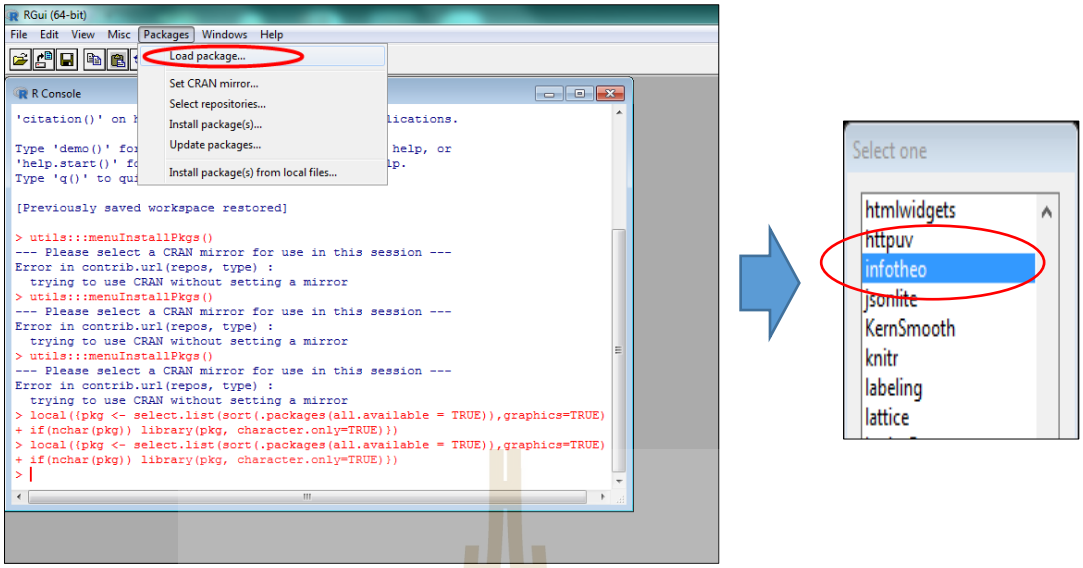
11) สิ้นสุดการทำงาน



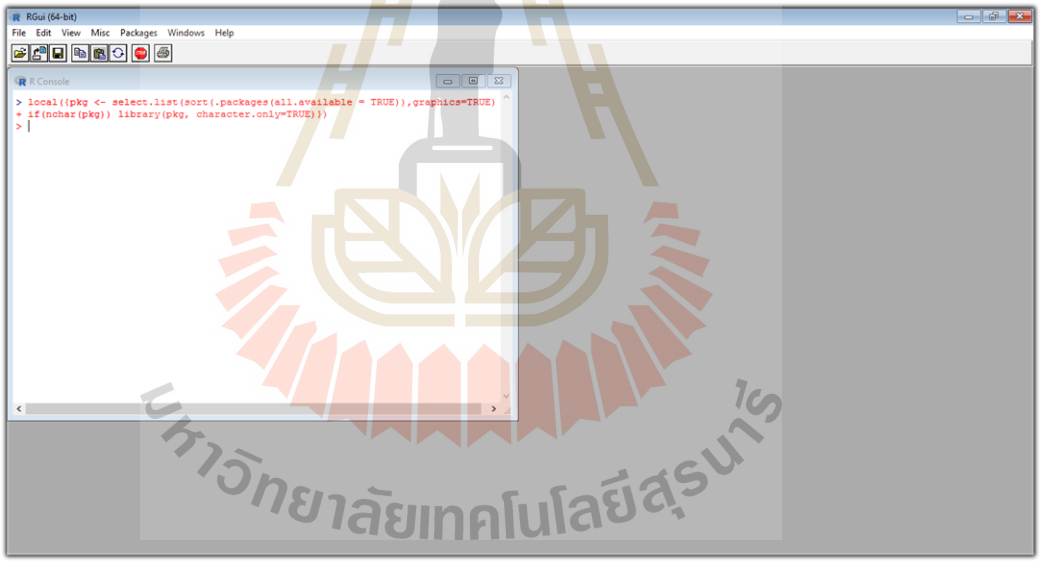
รูปที่ 3.9 หน้าต่างการใช้งานของโปรแกรม R



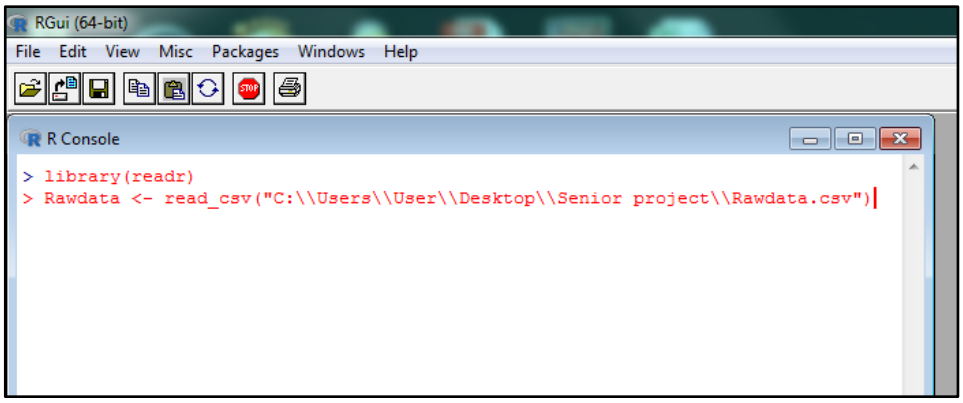
รูปที่ 3.10 วิธีติดตั้งแพ็กเกจ Infotheo



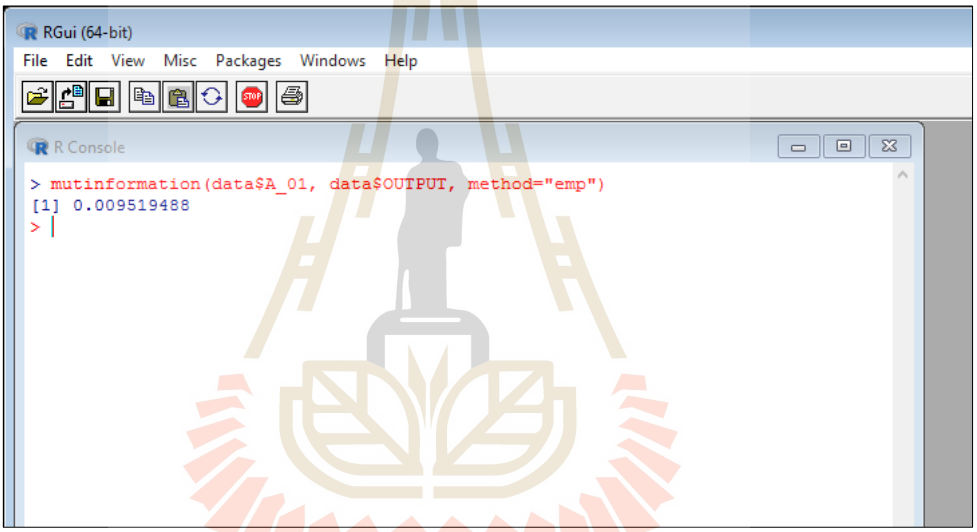
รูปที่ 3.11 วิธีโหลดแพ็กเกจ Infotheo



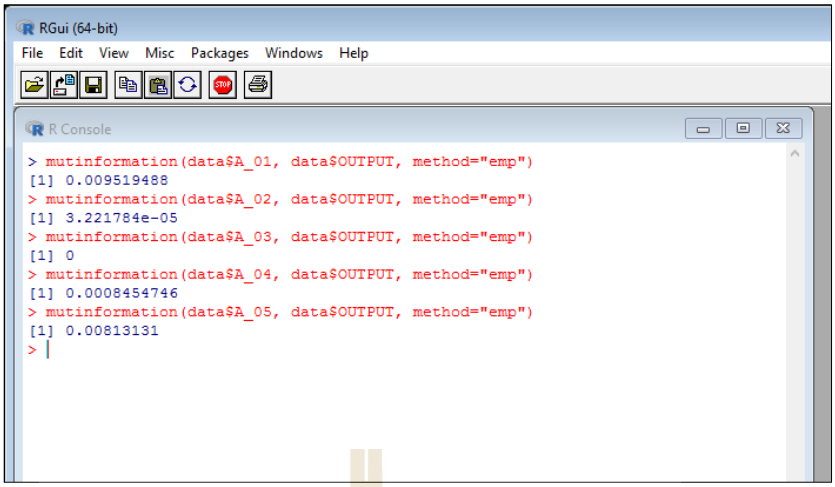
รูปที่ 3.12 หน้าต่างการใช้งานของแพ็คเกจ Infotheo



รูปที่ 3.13 วิธีนำเข้าข้อมูล excel เข้ามาในโปรแกรม R



รูปที่ 3.14 วิธีพิมพ์คำสั่งที่ใช้คำนวณ MI



รูปที่ 3.15 วิธีพิมพ์คำสั่งที่ใช้คำนวณ MI ในตัวแปรอิสระตัวอื่น ๆ

จากการดำเนินการวิเคราะห์ความสัมพันธ์โดยโปรแกรม R ด้วยวิธี Mutual Information ดังกล่าว จะเห็นได้ว่าข้อบกพร่องของแพ็กเกจ คือ ไม่สามารถคำนวณค่า MI ได้ครบทุกตัวแปรอิสระภายในทีเดียว ซึ่งการพิมพ์คำสั่งโดยผู้ใช้งานจนครบทุกชื่อตัวแปรอิสระทั้ง 74 ตัว จะใช้เวลาในการพิมพ์ข้อมูลประมาณ 2 ชั่วโมง ผู้วิจัยจึงทำการเขียนโค้ด VBA เพื่อวนลูปชื่อตัวแปรอิสระทั้ง 74 ตัว ซึ่งสามารถหาค่าได้ครบทุกตัวแปรอิสระภายในครั้งเดียว เมื่อผู้วิจัยเขียนโค้ดจาก VBA จากนั้นนำผลลัพธ์ที่ได้จากการเขียนโค้ดมาใส่ในโปรแกรม R ส่วนของการหาค่า MI ซึ่งการเขียนโปรแกรมในการวนลูปชื่อคอลัมน์ตัวแปรอิสระอธิบายดังด้านล่าง

Visual Basic มีชนิดของข้อมูลหลายชนิด ไม่ว่าจะเป็นตัวเลขจำนวนเต็ม ตัวเลขที่มีทศนิยม ข้อความ ตัวเลขทางการเงิน ค่าทางตรรกะ เป็นต้น ข้อมูลแต่ละชนิดจะใช้พื้นที่ในการเก็บไม่เท่ากัน ตัวแปรที่ใช้ในการเขียนโปรแกรมจะแตกต่างกันตามชนิดข้อมูล (data type) ใน VBA เช่น Integer, Single, String และVariant เป็นต้น (ชลาลัย วงเวียน, 2558) ในที่นี้ต้องการเขียนโปรแกรมให้วนลูปของชื่อคอลัมน์ของแต่ละตัวแปรอิสระให้ครบทุกตัวแปรและ ชื่อคอลัมน์ตัวแปรอิสระดังกล่าวเป็นจำนวนเต็ม จึงประกาศเป็นตัวแปร Integer ก่อนที่จะใช้งานตัวแปรหรือค่าคงที่ทุกครั้งควรประกาศตัวแปร (variable declaration) ก่อน เพื่อให้ Visual Basic รู้ว่าตัวแปรที่ต้องการใช้งานใช้แทนข้อมูลชนิดใดถึงแม้ว่า Visual Basic อนุญาตให้ใช้งานตัวแปรได้โดยไม่ต้องประกาศตัวแปร

Dim คือ คำสั่ง (statements) สำหรับประกาศตัวแปร

i คือ ชื่อของตัวแปรที่ต้องการประกาศ (ในที่นี้ให้ตัวแปร i นับตั้งแต่ A_01 ถึง A_09)

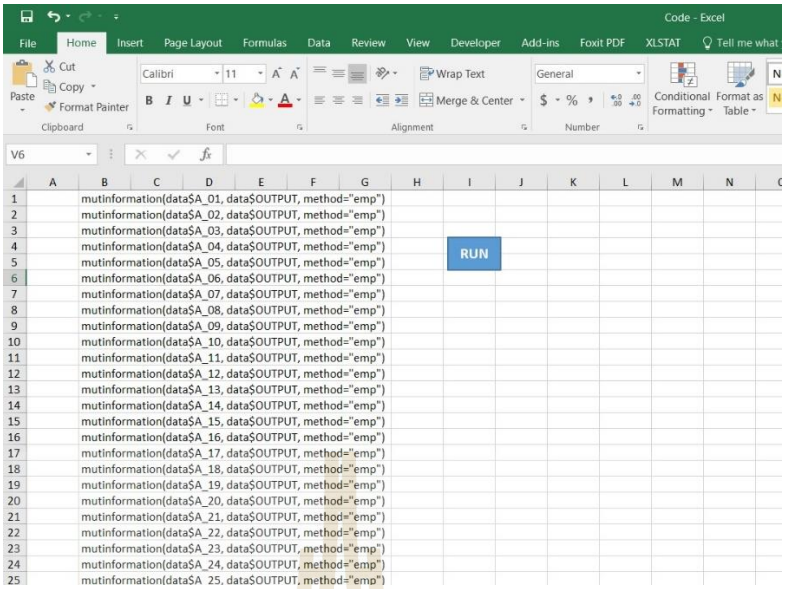
As คือ ส่วนที่บอกให้ Visual Basic ทราบว่าต้องการกำหนดชนิดของข้อมูล

Datatypes คือ ชนิดของข้อมูลที่ Visual Basic สนับสนุน ในที่นี้ data type คือ Integer

จากนั้นต้องการให้นับตัวแปร A_10 ถึง A_32 จึงประกาศตัวแปร j แทนตัวแปร i เนื่องจาก A_01 ถึง A_09 มีเลข 0 นำหน้าเลข 1-9 จึงต้องประกาศตัวแปรใหม่และ ประกาศตัวแปร k แทนตัวแปร j เนื่องจากเดิมเป็นตัวแปร Attribute (A) แต่ตัวแปรถัดไปคือ ตัวแปร Continuous (C) เปลี่ยนตัวแปรจนครบทุกชื่อคอลัมน์ของตัวแปรอิสระ เมื่อกดรันจะได้ผลลัพธ์จากการเขียนโปรแกรมด้วย VBA การเขียนโปรแกรม VBA และผลลัพธ์ที่ได้จากการรันแสดงดังรูปที่ 3.11 - 3.12 ตามลำดับ

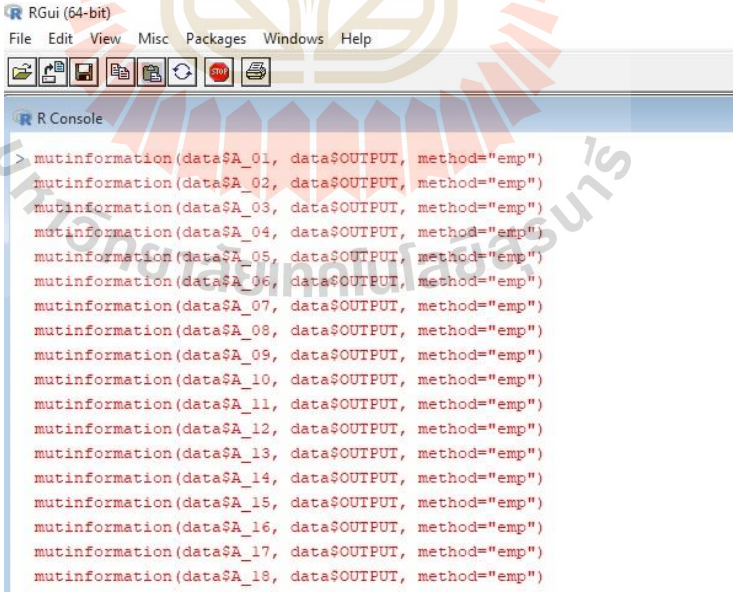
```
Sub MI()
Dim i As Integer
For i = 1 To 9
Cells(i, 2).Value = "mutinformation(data$A_0" & i & "" & ", data$OUTPUT, method=""emp""" & ")"

Next i

Dim j As Integer
For j = 10 To 32
Cells(j, 2).Value = "mutinformation(data$A_" & j & "" & ", data$OUTPUT, method=""emp""" & ")"
Next j

Dim k As Integer
For k = 33 To 38
Cells(k, 2).Value = "mutinformation(data$C_" & k & "" & ", data$OUTPUT, method=""emp""" & ")"
Next k

Dim l As Integer
For l = 39 To 46
Cells(l, 2).Value = "mutinformation(data$A_" & l & "" & ", data$OUTPUT, method=""emp""" & ")"
Next l

Dim m As Integer
For m = 47 To 60
Cells(m, 2).Value = "mutinformation(data$C_" & m & "" & ", data$OUTPUT, method=""emp""" & ")"
Next m

Dim n As Integer
For n = 61 To 65
Cells(n, 2).Value = "mutinformation(data$A_" & n & "" & ", data$OUTPUT, method=""emp""" & ")"
Next n

Dim o As Integer
For o = 66 To 68
Cells(o, 2).Value = "mutinformation(data$C_" & o & "" & ", data$OUTPUT, method=""emp""" & ")"
Next o

Dim p As Integer
For p = 69 To 74
Cells(p, 2).Value = "mutinformation(data$A_" & p & "" & ", data$OUTPUT, method=""emp""" & ")"
Next p

End Sub
```

รูปที่ 3.16 แสดงโค้ดจากการเขียนด้วย VBA บนโปรแกรม Excel

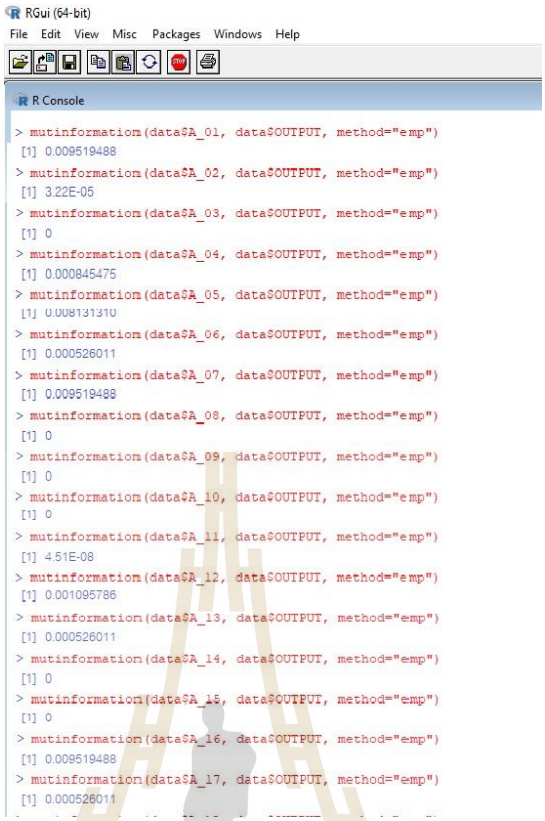


รูปที่ 3.17 แสดงผลที่ได้ส่วนหนึ่งจากการเขียนโค้ด VBA บนโปรแกรม Excel

เมื่อทำการรันผลลัพธ์ที่ได้จากการเขียนโค้ด VBA จะได้ผลลัพธ์ดังรูปที่ 3.17 จากนั้นนำผลลัพธ์ที่ได้ไปใส่ในส่วนของการหา MI ในโปรแกรม R ดังขั้นตอนที่ 8) จะได้ค่า MI Value



รูปที่ 3.18 แสดงการหน้าต่างของโปรแกรม R หลังจากทำการคัดลอกผลลัพธ์จาก VBA มาใส่

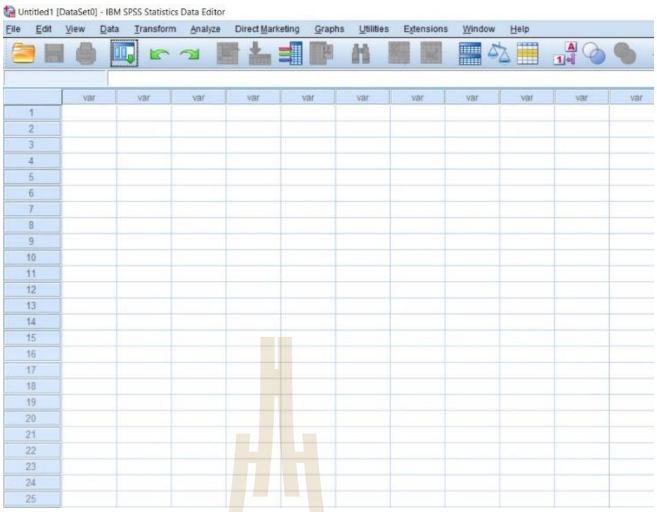


รูปที่ 3.19 แสดงค่า MI Value ทั้งหมด

3.5.3 ขั้นตอนการทวนสอบผลโดยโปรแกรม SPSS ด้วยวิธี Logistic Regression

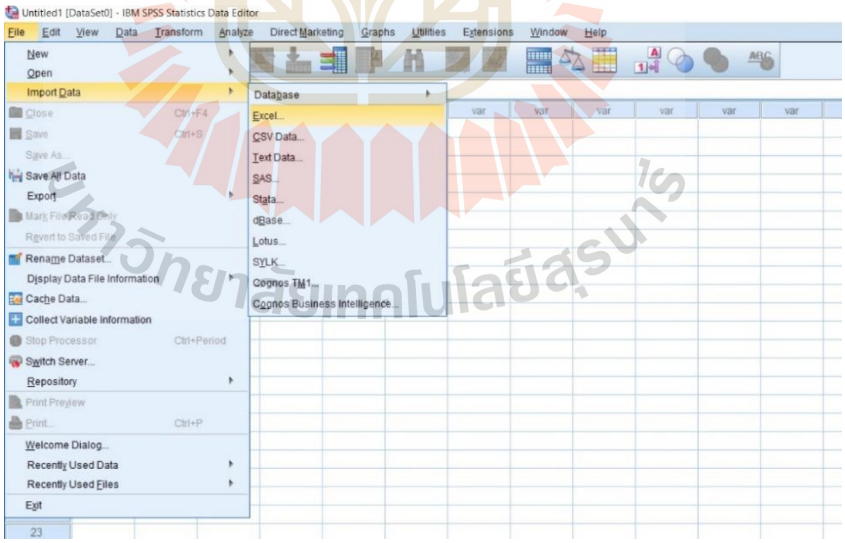
จากการวิเคราะห์ข้อมูลการวิจัยซึ่งเหมาะกับวิธีการถดถอยโลจิสติกส์ในตารางที่ 3.1 โดยตัวแปรตามคือ คุณภาพของฮาร์ดดิสก์มีค่าเท่ากับ 0 หรือ 1 (Binary Variable) เท่านั้นซึ่งเป็นตัวแปรเชิงคุณภาพและ ตัวแปรอิสระคือ ตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ซึ่งมีทั้งหมด 58 ตัวแปร มีค่าเป็นทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ (Continuous Variables and Attribute Variables) โดยขั้นตอนการหาค่าด้วยวิธี Logistic Regression แสดงดังขั้นตอนที่ 1) – 9)

1) เปิดโปรแกรม SPSS (ดังรูปที่ 3.20)



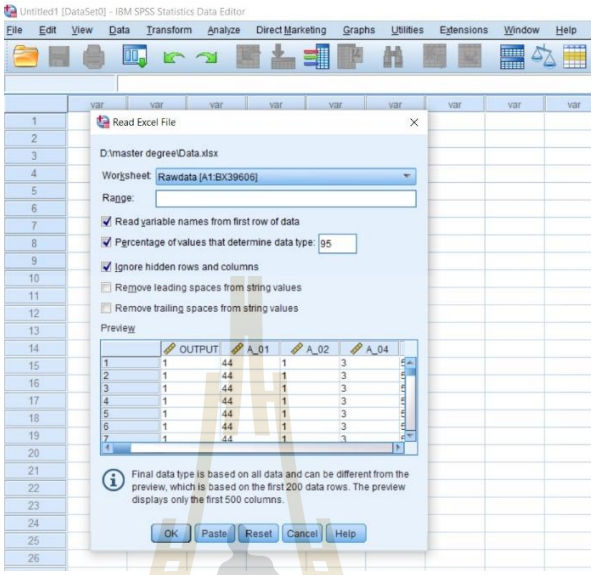
รูปที่ 3.20 แสดงหน้าต่างการใช้งานของโปรแกรม SPSS

2) ทำการนำเข้าข้อมูลจากไฟล์ Excel เข้ามาในโปรแกรม SPSS โดยเข้าไปที่ File > Import Data > Excel (ดังรูปที่ 3.21)



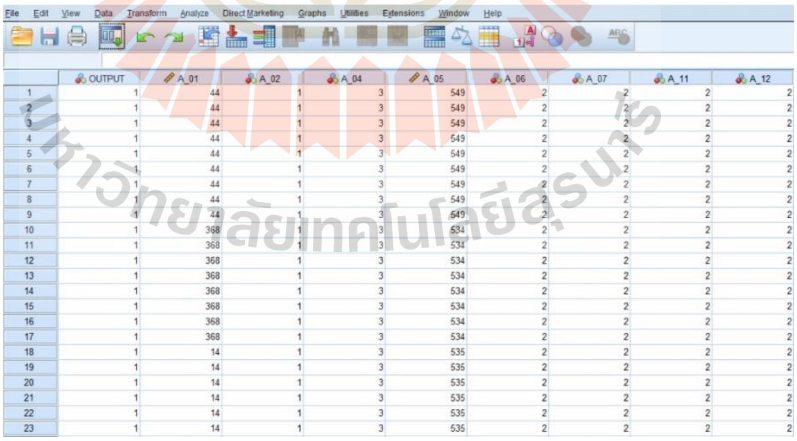
รูปที่ 3.21 แสดงวิธีการนำเข้าข้อมูลจากไฟล์ Excel เข้ามาในโปรแกรม SPSS

3) ทำการคลิกเลือกไฟล์ข้อมูล Excel ที่ต้องการคำนวณเข้ามายังโปรแกรม โดยโปรแกรมจะแสดงหน้าตาของข้อมูลในการคำนวณ จากนั้นคลิก OK (ดังรูปที่ 3.22)



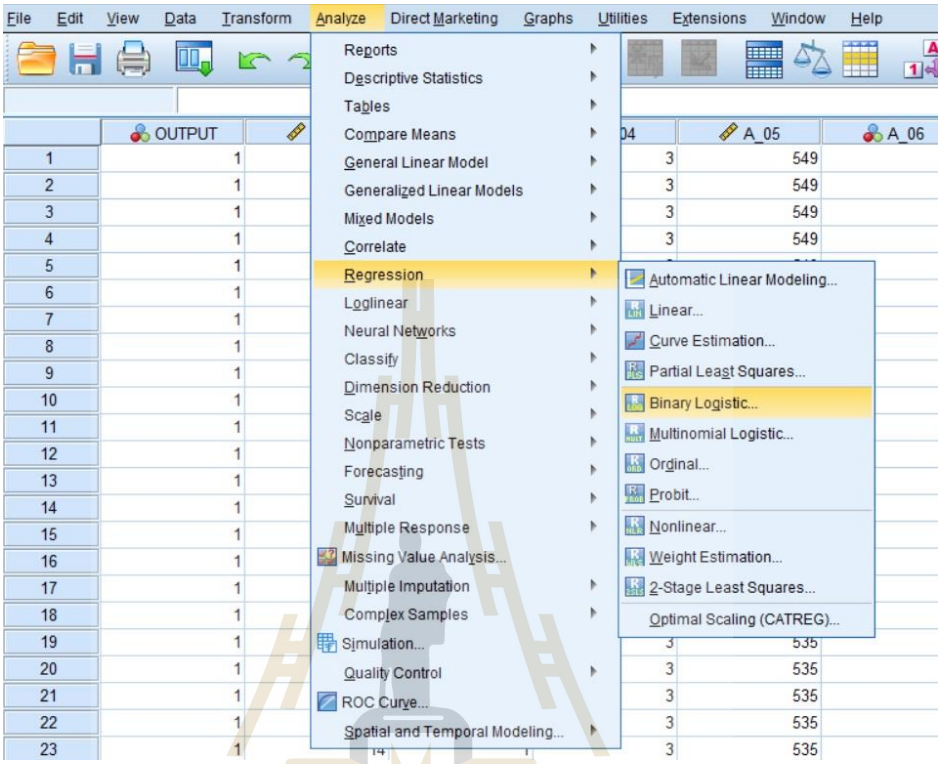
รูปที่ 3.22 แสดงไฟล์ข้อมูล Excel ที่ทำการเลือกเข้ามายังโปรแกรม SPSS

4) โปรแกรมจะทำการนำเข้าข้อมูลจากไฟล์ที่เลือก (ดังรูปที่ 3.23)



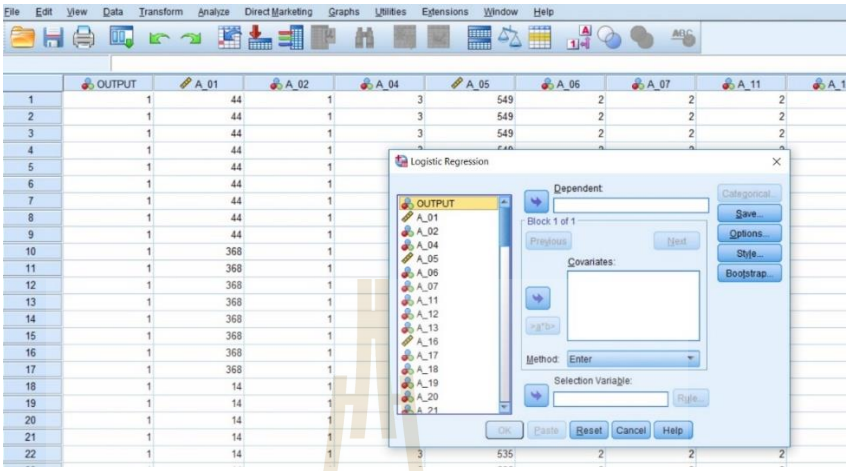
รูปที่ 3.23 แสดงหน้าต่างของโปรแกรมหลังจากทำการนำเข้าข้อมูล

5) ทำการหาค่าความสัมพันธ์ของข้อมูลด้วยวิธี Binary Logistic Regression โดยเข้าไปที่ Analyze > Regression > Binary Logistic (ดังรูปที่ 3.24)



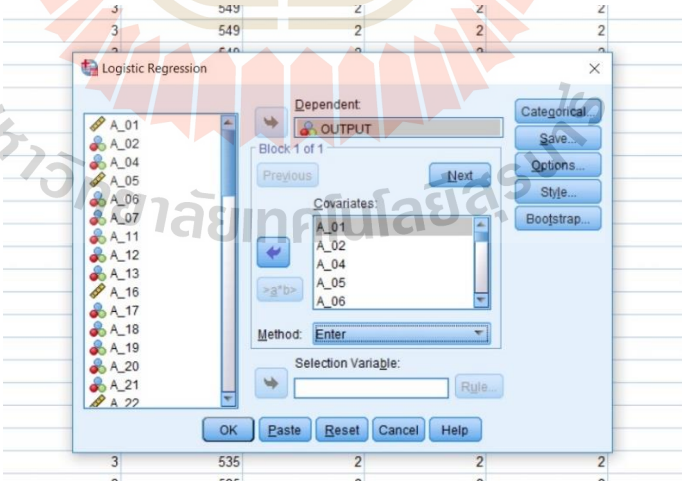
รูปที่ 3.24 แสดงวิธีการหาค่าความสัมพันธ์ของข้อมูลด้วยวิธี Binary Logistic Regression

6) โปรแกรมจะขึ้นหน้าต่างของการคำนวณ Logistic Regression โดยให้ผู้ใช้งานเลือกป้อนตัวแปรแต่ละตัวที่ใช้ในการคำนวณ (ดังรูปที่ 3.25)



รูปที่ 3.25 แสดงหน้าต่างของการคำนวณ Logistic Regression

7) ในส่วนของแถบ Dependent คลิกเลือกตัวแปรตาม คือตัวแปรที่ชื่อว่า OUTPUT และส่วนของ Covariates คลิกเลือกตัวแปรอิสระที่มีผลต่อคุณภาพของฮาร์ดดิสก์ทั้ง 58 ตัวแปร จากนั้นกด OK (ดังรูปที่ 3.26)



รูปที่ 3.26 แสดงการเลือกตัวแปรตามและตัวแปรอิสระในการคำนวณ

8) ผลลัพธ์ที่ได้จะแสดงในหน้า OUTPUT ซึ่งอยู่อีกหน้าต่างหนึ่งของโปรแกรม R

			Score
Step 0	Variables	A_01	33.994
		A_02	.304
		A_04	15.766
		A_05	53.913
		A_06	24.709
		A_07	39605.013
		A_11	.004
		A_12	50.410
		A_13	25.969
		A_16	1.162
		A_17	25.969
		A_18	35794.883
		A_19	39604.987
		A_20	39605.013
		A_21	39604.987
		A_22	1.011
		A_23	.318
		A_24	12.826
		A_25	28.392
		A_26	6.530
		A_27	.005

รูปที่ 3.27 ผลลัพธ์ส่วนหนึ่งจากการหาค่าความสัมพันธ์ด้วยวิธี Logistic Regression

9) สิ้นสุดการทำงาน

บทที่ 4
ผลการดำเนินการวิจัย

ในบทนี้นำเสนอผลการดำเนินการวิจัยที่ได้จากการวิจัยครั้งนี้ มีการนำเสนอผลการวิเคราะห์ในแต่ละหัวข้อซึ่งแบ่งออกเป็น 1) ผลการดำเนินการวิจัยจากการหาค่าความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม R ด้วยวิธี Mutual Information และ 2) ผลการทวนสอบความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม SPSS® ด้วยวิธี Logistic Regression

4.1 ผลการดำเนินการวิจัยจากการหาค่าความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม R ด้วยวิธี Mutual Information

จากการดำเนินการวิจัยหาค่าความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม R ด้วยวิธี Mutual Information ได้ MI Value แสดงดังตารางที่ 4.1

ตารางที่ 4.1 แสดงค่า MI จากการคำนวณด้วยวิธี Mutual Information

Variable	MI Value
A_01	0.009519488
A_02	3.22E - 05
A_03	0
A_04	0.000845475
A_05	0.008131310
A_06	0.000526011
A_07	0.009519488
A_08	0
A_09	0
A_10	0
A_11	4.51E - 08
A_12	0.001095786

ตารางที่ 4.1 แสดงค่า MI จากการคำนวณด้วยวิธี Mutual Information (ต่อ)

Variable	MI Value
A_13	0.000526011
A_14	0
A_15	0
A_16	0.009519488
A_17	0.000526011
A_18	0.009519488
A_19	0.009519488
A_20	0.009519488
A_21	0.009519488
A_22	0.009519488
A_23	0.000645846
A_24	0.001328113
A_25	0.000892669
A_26	0.00304748
A_27	0.000780168
A_28	0.009519488
A_29	0
A_30	0
A_31	0
A_32	0
A_39	0.00359721
A_40	0.00052668
A_41	0
A_42	0.000938156
A_43	0.003444311
A_44	0
A_45	0

ตารางที่ 4.1 แสดงค่า MI จากการคำนวณด้วยวิธี Mutual Information (ต่อ)

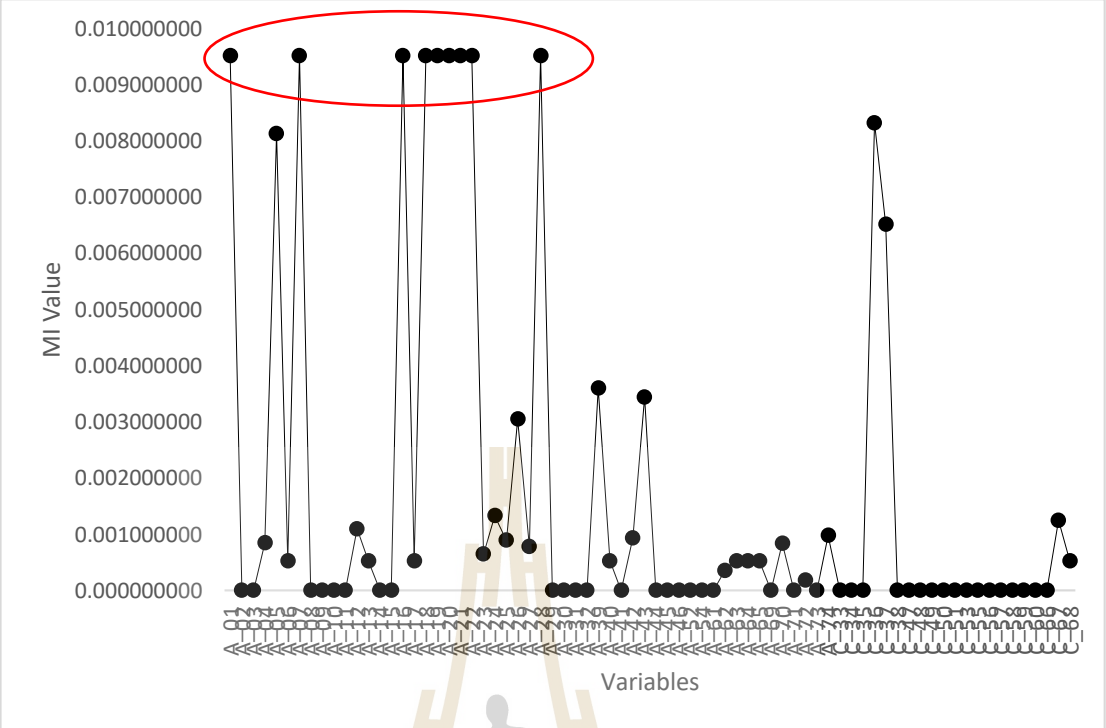
Variable	MI Value
A_46	0
A_52	0
A_54	0
A_61	0
A_62	0.000351165
A_63	0.000526011
A_64	0.000526011
A_65	0.000526011
A_69	3.53E - 05
A_70	0.000843444
A_71	0
A_72	0.000183608
A_73	2.13E - 05
A_74	0.00097782
C_33	0
C_34	0
C_35	0
C_36	0.008324255
C_37	0.006514658
C_38	0
C_47	0
C_48	0
C_49	0
C_50	0
C_51	0
C_53	0
C_55	0

ตารางที่ 4.1 แสดงค่า MI จากการคำนวณด้วยวิธี Mutual Information (ต่อ)

Variable	MI Value
C_56	0
C_57	2.52E - 07
C_58	0
C_59	0
C_60	0
C_66	0
C_67	0.001249962
C_68	0.000526011

4.1.1 การวิเคราะห์ผลจากการคำนวณด้วยวิธี Mutual Information

จากการวิเคราะห์ตารางที่ 4.1 เป็นการหาค่าความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อคุณภาพของฮาร์ดดิสก์โดยใช้โปรแกรม R ด้วยวิธี Mutual Information คือ เมื่อค่าของ MI มีค่าเท่ากับ 0จะถือว่าไม่มีความสัมพันธ์กันของข้อมูล แต่ถ้าหากค่า MI ของข้อมูลมีค่ามากที่สุดจะถือได้ว่าข้อมูลมีความสัมพันธ์มากที่สุด จากนั้นนำค่า MI ที่ได้จากการคำนวณมาพล็อตกราฟเพื่อดูค่า MI ที่มีค่ามาก



รูปที่ 4.1 กราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ MI Value

จากรูปที่ 4.1 เมื่อทำการพล็อตกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ MI Value จะเห็นได้ว่าค่า MI Value ที่มีค่าสูงคือ ตัวแปร A_01, A_07, A_16, A_18, A_19, A_20, A_21, A_22 และ A_28

4.2 ผลการทวนสอบความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม SPSS® ด้วยวิธี Logistic Regression

จากการดำเนินการวิจัยหาค่าความสัมพันธ์ของข้อมูลโดยใช้โปรแกรม SPSS® ด้วยวิธี Logistic Regression เพื่อทวนสอบผล ได้ Score แสดงดังตารางที่ 4.2

ตารางที่ 4.2 แสดง Score จากการคำนวณด้วยวิธี Logistic Regression

Variable	Score
A_01	33.99429
A_02	0.30426
A_04	15.76633
A_05	53.91322
A_06	24.70874
A_07	39605.01
A_11	0.003508
A_12	50.40973
A_13	25.96921
A_16	1.161705
A_17	25.96916
A_18	35794.88
A_19	39604.99
A_20	39605.01
A_21	39604.99
A_22	1.011486
A_23	0.317657
A_24	12.82611
A_25	28.39208
A_26	6.529822
A_27	0.005457
A_28	39605.02

ตารางที่ 4.2 แสดง Score จากการคำนวณด้วยวิธี Logistic Regression (ต่อ)

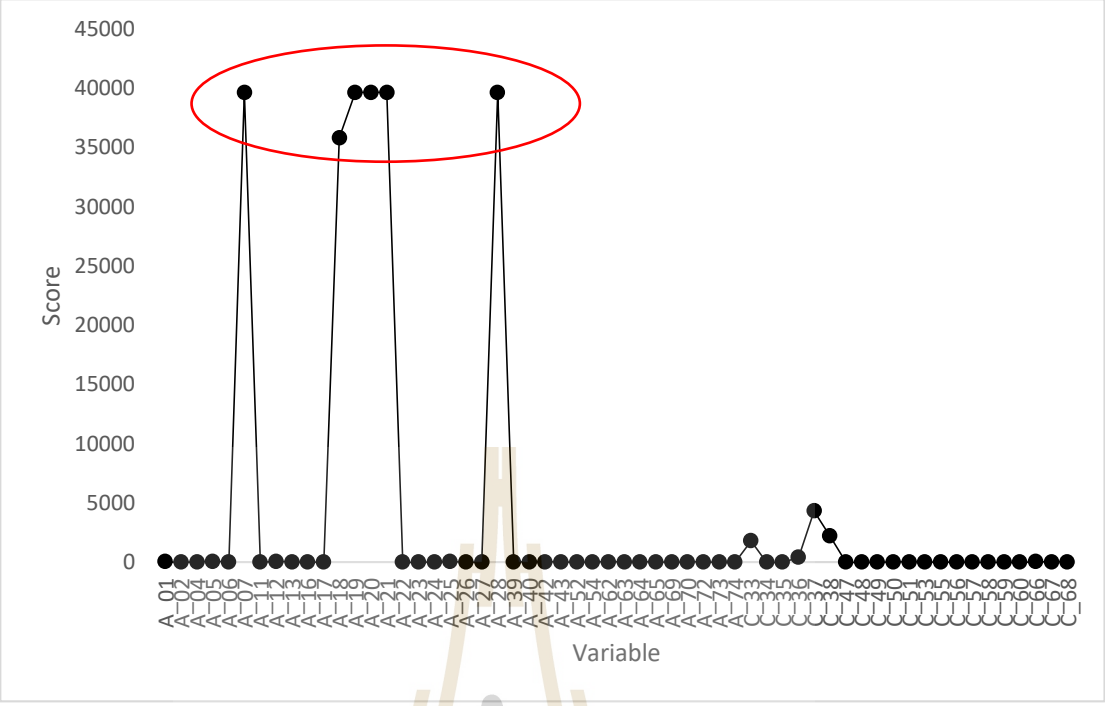
Variable	Score
A_39	9.493716
A_40	25.4476
A_42	23.05433
A_43	26.31965
A_52	0.060753
A_54	18.01042
A_62	25.07876
A_63	25.96922
A_64	25.96922
A_65	16.59341
A_69	1.543499
A_70	25.52092
A_72	5.508935
A_73	1.469338
A_74	0.322971
C_33	1804.469
C_34	2.123813
C_35	3.369443
C_36	408.4829
C_37	4322.44
C_38	2202.393
C_47	0.522372
C_48	4.185113
C_49	0.548506
C_50	5.836463
C_51	13.3804
C_53	15.36177

ตารางที่ 4.2 แสดง Score จากการคำนวณด้วยวิธี Logistic Regression (ต่อ)

Variable	Score
C_55	0.47309
C_56	23.93921
C_57	0.019003
C_58	1.026098
C_59	0.97152
C_60	2.232858
C_66	41.913
C_67	7.213522
C_68	19.01106

4.2.1 การวิเคราะห์ผลการทวนสอบจากการคำนวณด้วยวิธี Logistic Regression

จากการวิเคราะห์ตารางที่ 4.2 เป็นการผลการทวนสอบความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อคุณภาพของฮาร์ดดิสก์โดยใช้โปรแกรม SPSS ด้วยวิธี Logistic Regression คือ เมื่อ Score มีค่าเท่ากับ 0 จะถือว่าไม่มีความสัมพันธ์กันของข้อมูล แต่ถ้าหากค่า Score ของข้อมูลมีค่ามากที่สุดจะถือได้ว่าข้อมูลมีความสัมพันธ์มากที่สุด จากนั้นนำ Score ที่ได้จากการคำนวณมาพล็อตกราฟเพื่อดู Score ที่มีค่ามาก



รูปที่ 4.2 กราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ Score

จากรูปที่ 4.2 เมื่อทำการพล็อตกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ Score จะเห็นได้ว่า Score ที่มีค่าสูงคือ ตัวแปร A_07, A_18, A_19, A_20, A_21 และ A_28

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษา

การศึกษาวิจัยนี้มีวัตถุประสงค์คือ 1. เพื่อหาค่าความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ที่มีผลต่อการทดสอบคุณภาพของฮาร์ดดิสก์และ 2. เพื่อศึกษาตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ของบริษัทกรณีศึกษา โดยข้อมูลที่ได้จากบริษัทกรณีศึกษาวัดค่าจากกระบวนการตรวจสอบคุณภาพของฮาร์ดดิสก์ ข้อมูลที่ได้รับมีจำนวนตัวแปรคุณภาพ (ตัวแปรตาม) ของฮาร์ดดิสก์ทั้งหมด 57,232 ตัว และจำนวนตัวแปรที่เกี่ยวข้องต่อคุณภาพของฮาร์ดดิสก์ (ตัวแปรอิสระ) 74 ตัว ซึ่งแบ่งเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจำนวน 53 ตัวและ ตัวแปรเชิงปริมาณจำนวน 21 ตัว ก่อนทำการวิจัยผู้วิจัยได้ทำการลบแถวของคุณภาพฮาร์ดดิสก์ที่ไม่มีผลคุณภาพของฮาร์ดดิสก์ว่าผ่านหรือไม่ผ่านการตรวจสอบคุณภาพและ คุณภาพของฮาร์ดดิสก์ที่มีตัวแปรไม่ครบทั้ง 74 ตัวแปรออก ซึ่งจะเหลือข้อมูลจำนวนตัวแปรคุณภาพของฮาร์ดดิสก์ในการคำนวณ 39,605 ตัวและ จำนวนตัวแปรที่เกี่ยวข้องต่อคุณภาพของฮาร์ดดิสก์ 74 ตัวดังเดิม ซึ่งแบ่งเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจำนวน 53 ตัว และ ตัวแปรเชิงปริมาณจำนวน 21 ตัว

การวิจัยนี้ใช้การหาค่าความสัมพันธ์ของข้อมูลด้วยวิธี Mutual Information โดยใช้โปรแกรม R และใช้วิธี Logistic Regression โดยใช้โปรแกรม SPSS® ในการทวนสอบผล จากการหาค่าความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อคุณภาพของฮาร์ดดิสก์โดยใช้โปรแกรม R ด้วยวิธี Mutual Information คือ เมื่อ MI Vale มีค่าเท่ากับ 0 จะถือว่าไม่มีความสัมพันธ์กันของข้อมูล แต่ถ้าหากค่า MI ของข้อมูลมีค่ามากที่สุดจะถือได้ว่าข้อมูลมีความสัมพันธ์มากที่สุดและ จากการทวนสอบความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อคุณภาพของฮาร์ดดิสก์โดยใช้โปรแกรม SPSS® ด้วยวิธี Logistic Regression คือ เมื่อ Score มีค่าเท่ากับ 0 จะถือว่าไม่มีความสัมพันธ์กันของข้อมูล แต่ถ้าหากค่า Score ของข้อมูลมีค่ามากที่สุดจะถือได้ว่าข้อมูลมีความสัมพันธ์มากที่สุด จากนั้นทำการพล็อตกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ MI Value จะเห็นได้ว่าค่า MI Value ที่มีค่าสูงคือ ตัวแปร A_01, A_07, A_16, A_18, A_19, A_20, A_21, A_22 และ A_28 และทำการพล็อตกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระต่าง ๆ กับ Score จะเห็นได้ว่า Score ที่มีค่าสูงคือ ตัวแปร A_07, A_18, A_19, A_20, A_21 และ A_28

ตารางที่ 5.1 การเรียงลำดับค่าความสัมพันธ์ของแต่ละวิธี

ลำดับที่	Mutual Information		Logistic Regression	
1	A_01	0.009519488	A_28	39605.021
2	A_07	0.009519488	A_20	39605.013
3	A_16	0.009519488	A_07	39605.013
4	A_18	0.009519488	A_21	39604.987
5	A_19	0.009519488	A_19	39604.987
6	A_20	0.009519488	A_18	35794.883
7	A_21	0.009519488	C_37	4322.440
8	A_22	0.009519488	C_38	2202.393
9	A_28	0.009519488	C_33	1804.469

จากตารางที่ 5.1 เมื่อทำการเรียงลำดับค่ามากที่สุดของแต่ละวิธี จะเห็นได้ว่าตัวแปรอิสระที่ซ้ำกันคือ ตัวแปร A_07, A_18, A_19, A_20, A_21 และ A_28 กล่าวคือ ตัวแปรอิสระดังกล่าวส่งผลต่อคุณภาพของฮาร์ดดิสก์ในกระบวนการทดสอบ

โดยสรุปแล้วงานวิจัยนี้บรรลุวัตถุประสงค์ดังกล่าวแล้วในบทที่ 1 หัวข้อที่ 1.2

5.2 ข้อสังเกต

เมื่อได้ผลสรุปของตัวแปรที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์จากตารางที่ 5.1 แล้ว ผู้วิจัยได้ทำการกรองข้อมูลของตัวแปรแต่ละตัว ซึ่งลักษณะของข้อมูลคือ ตัวแปร A_07 มีค่าของตัวแปรเพียง 2 ค่านั่นคือ 1 และ 2 เมื่อทำการเลือกค่าตัวแปร A_07 มีค่าเท่ากับ 1 จะส่งผลให้คุณภาพของฮาร์ดดิสก์มีค่าเป็น 0 กล่าวคือ ไม่ผ่านกระบวนการทดสอบคุณภาพ แต่หากตัวแปร A_07 มีค่าเท่ากับ 2 จะส่งผลให้คุณภาพของฮาร์ดดิสก์มีค่าเป็น 1 กล่าวคือ ผ่านกระบวนการทดสอบคุณภาพ ในขณะเดียวกันตัวแปร A_18 มีค่าของตัวแปรเท่ากับ 1, 2 และ 3 เมื่อทำการเลือกค่าตัวแปร A_18 มีค่าเท่ากับ 2 หรือ 3 จะส่งผลให้คุณภาพของฮาร์ดดิสก์มีค่าเป็น 0 กล่าวคือ ไม่ผ่านกระบวนการทดสอบคุณภาพ แต่หากตัวแปร A_18 มีค่าเท่ากับ 1 จะส่งผลให้คุณภาพของฮาร์ดดิสก์มีค่าเป็น 1 กล่าวคือ ผ่านกระบวนการทดสอบคุณภาพ ซึ่งสามารถสรุปดังตารางที่ 5.2

ตารางที่ 5.2 ผลสรุปเมื่อทำการเปลี่ยนค่าของตัวแปรอิสระ ซึ่งจะส่งผลต่อคุณภาพของฮาร์ดดิสก์

ตัวแปรอิสระ	ค่าของตัวแปร	คุณภาพของฮาร์ดดิสก์
A_07	1	ไม่ผ่านกระบวนการทดสอบคุณภาพ
	2	ผ่านกระบวนการทดสอบคุณภาพ
A_18	1	ผ่านกระบวนการทดสอบคุณภาพ
	2	ไม่ผ่านกระบวนการทดสอบคุณภาพ
	3	ไม่ผ่านกระบวนการทดสอบคุณภาพ
A_19	1	ผ่านกระบวนการทดสอบคุณภาพ
	2	ไม่ผ่านกระบวนการทดสอบคุณภาพ
A_20	1	ไม่ผ่านกระบวนการทดสอบคุณภาพ
	2	ผ่านกระบวนการทดสอบคุณภาพ
A_21	1	ผ่านกระบวนการทดสอบคุณภาพ
	2	ไม่ผ่านกระบวนการทดสอบคุณภาพ
A_28	2	ไม่ผ่านกระบวนการทดสอบคุณภาพ
	3	ผ่านกระบวนการทดสอบคุณภาพ

5.3 ข้อเสนอแนะ

1) งานวิจัยนี้ศึกษาค่าความสัมพันธ์ของตัวแปรต่าง ๆ ที่ส่งผลต่อคุณภาพของฮาร์ดดิสก์ งานวิจัยต่อไปควรมีการศึกษาควบคู่กับการพยากรณ์คุณภาพของฮาร์ดดิสก์ เพื่อเช็คความแม่นยำว่าตัวแปรอิสระที่คำนวณได้ส่งผลต่อคุณภาพของฮาร์ดดิสก์มากน้อยเพียงใด

2) วิธีการที่นำเสนอในงานวิจัยนี้ มีรูปแบบการใช้งานง่าย แต่ควรพิจารณาเงื่อนไขและ ข้อจำกัดต่าง ๆ ของข้อมูลที่จะนำมาใช้

3) โปรแกรม R เป็นโปรแกรมที่เหมาะแก่การวิเคราะห์ข้อมูลมากกว่าโปรแกรม SPSS® เนื่องจากเป็นโปรแกรม Open Source Software สามารถใช้ได้ฟรี แต่มีความยุ่งยากในการเขียนโค้ดโปรแกรม หากผู้ใช้ไม่มีความรู้ด้านการเขียนภาษา R

รายการอ้างอิง

กัลยา วานิชย์บัญชา (2549). **สถิติสำหรับงานวิจัย.** พิมพ์ครั้งที่ 2.

กัลยา วานิชย์บัญชา (2550). **การวิเคราะห์ข้อมูลหลายตัวแปร.** พิมพ์ครั้งที่ 2.

กัลยา วานิชย์บัญชา (2554). **การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows.** พิมพ์ครั้งที่ 9.

กัลยา วานิชย์บัญชา (2548). **การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล.** พิมพ์ครั้งที่ 7.

กัลยา วานิชย์บัญชา (2546). **การวิเคราะห์สถิติ : สถิติสำหรับการบริหารและวิจัย.** พิมพ์ครั้งที่ 14.

จามรี ชูบัวทอง และสมศรี บัณฑิตวิไล (2560). การพัฒนาตัวแบบเพื่อพยากรณ์คุณภาพผลิตภัณฑ์ฮาร์ดดิสก์ด้วยการถดถอยโลจิสติกส์และโครงข่ายประสาทเทียมโดยใช้การวิเคราะห์เหมืองข้อมูล. **วารสารวิทยาศาสตร์และเทคโนโลยี.** ปีที่ 25 ฉบับที่ 1. 1-13.

ชลาลัย วงเวียน (2558). การสร้างแผนภูมิความคลาดเคลื่อนด้วยโปรแกรมตารางจัดการ. **ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมอุตสาหการ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.**

ชูศรี วงศ์รัตนะ (2560). **เทคนิคการใช้สถิติเพื่อการวิจัย.** พิมพ์ครั้งที่ 13. กรุงเทพฯ: อมรการพิมพ์.

ณาตยา แดงรุ่งโรจน์ (2014). **ความสามารถในการแข่งขันส่งออกฮาร์ดดิสก์ไดร์ฟของประเทศไทยกับประเทศสมาชิกอาเซียน 4 ประเทศ**. สุทธิปริทัศน์ ปีที่ 28 ฉบับที่ 85. 161-185.

ธนดล สุชาตพงศ์ (2557). การตรวจสอบคุณภาพของฮาร์ดดิสก์ภายใต้มาตรฐานอุตสาหกรรม ด้วยวิธีการเหมืองข้อมูล. **ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี.**

ธีรนพ สุขอารมณ์ และ ปารเมศ ชุติมา (2556). การปรับปรุงกระบวนการทดสอบฮาร์ดดิสก์ไดร์ฟเพื่อลดข้อ บกพร่องประเภทการอ่าน/เขียนสัญญาณบกพร่องของหัวอ่าน/เขียน. **Engineering Journal of Research and Development**. 24(1). 67-74.

นิรมล พันสีมาและ อนัตต์ เจ่าสกุล (2557). การเปรียบเทียบการทำงานโปรแกรม R และโปรแกรม SPSS กรณีการจำแนกประเภทข้อมูลเงินยืมทดรองจ่ายของมหาวิทยาลัยขอนแก่นภายใต้แนวความคิดการทำเหมืองข้อมูล. **วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอุบลราชธานี.** ปีที่ 16 ฉบับที่ 1.

ประสาน นาคอ่อน (2557). กลยุทธ์ในการปรับปรุงประสิทธิภาพของกระบวนการทดสอบฮาร์ดดิสก์ไดรฟ์. **ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาแมคคาทรอนิกส์ มหาวิทยาลัยเทคโนโลยีสุรนารี.**

ภัทรศยา ตันติวัฒนกูล และ อรรถกร เก่งพล (2556). การพยากรณ์มูลค่าต้นทุนของเสียในอุตสาหกรรมการผลิตฮาร์ดดิสก์ไดรฟ์. **วารสารวิชาการพระจอมเกล้าพระนครเหนือ.** ปีที่ 23 ฉบับที่ 1. 148-156.

วราฤทธิ์ พานิชกิจโกศลกุล (2550). **การใช้โปรแกรม R ในงานวิจัยด้านทฤษฎีสถิติ**. ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์.

วิโรจน์ อรุณมานะกุล (2559). **สถิติและการใช้โปรแกรม R**. ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย พงษ์วิชัย. (2556) **การวิเคราะห์ข้อมูลทางสถิติด้วยคอมพิวเตอร์ เน้นสำหรับงานวิจัย.** พิมพ์ครั้งที่ 24. กรุงเทพฯ : สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย

ศูนย์วิจัยกสิกรไทย (2558). สืบค้นจาก http://www.sereechai.com/index.php/2013-05-01-06-35-10/2013-05-01-07-33-49/3336-2015-04-23-22-37-36

สมจิต วัฒนาชยากูล (2532). **สถิติวิเคราะห์เบื้องต้น.** พิมพ์ครั้งที่ 3. กรุงเทพฯ: สำนักพิมพ์ประกายพรึก.

อโณทัย ศิลเทพาเวทย์ (2554). **แบบจำลองเพื่อพัฒนาคุณภาพของผลิตภัณฑ์เอชจีเอในโรงงานอุตสาหกรรมฮาร์ดดิสก์ด้วยเทคนิคต้นไม้ตัดสินใจ**. จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

Haodi Jiang, Turki Turki and Jason T. L. Wang (2017). Reverse Engineering Regulatory Networks in Cells Using a Dynamic Bayesian Network and Mutual Information Scoring Function. **IEEE International Conference on Machine Learning and Applications.** DOI 10.1109/ICMLA. 00-67.

Luis M. de Campos (2006). A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests. **Journal of Machine Learning Research 7.** 2149-2187.

Murray, J. F., Hughes, G. F., & Kreutz-Delgado, K (2005). Machine learning methods for predicting failures in hard drives: A multiple-instance application. **Journal of Machine Learning Research**. 6(May). 783-816.

Nara Samattapapong (2559). A production Throughput Forecasting System in an Automated Hard Disk Drive Test Operation Using GRNN. **Journal of Industrial Engineering and Management.** Vol 9. No.2 . 330-358.

Pinheiro, E., Weber, W. D., & Barroso, L. A. (2007, February). **Failure Trends in a Large Disk Drive Population.** In FAST Vol. 7 No. 1. 17-23.

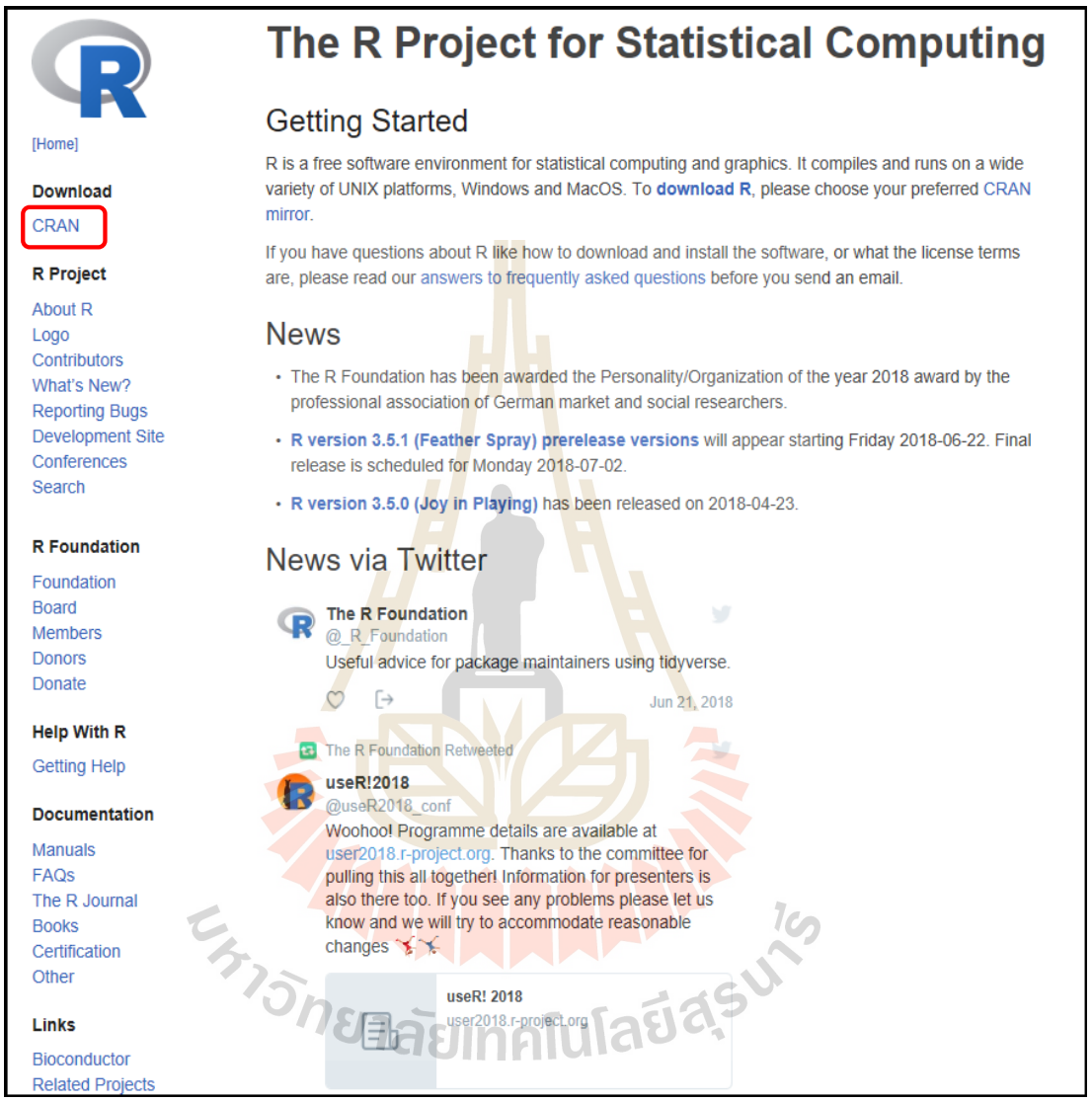
Shannon, C.E. (1948). A mathematical theory of communication**. Bell System Technical Journal.** 27(3). 379-423.

ภาคผนวก ก

วิธีการติดตั้งโปรแกรม R

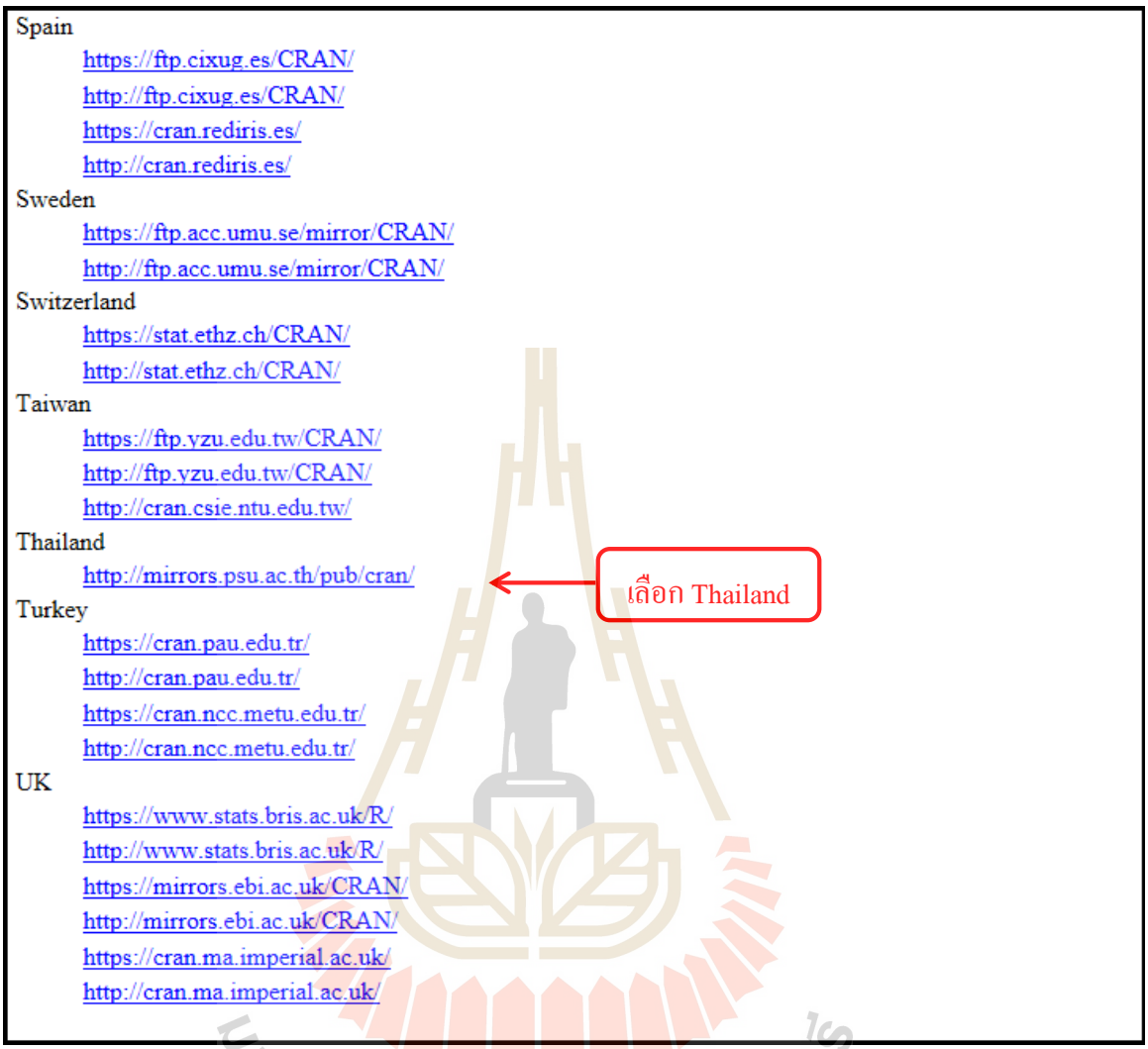
วิธีการติดตั้งโปรแกรม R

1) เปิดเว็บไซต์ http://www.r-project.org/ แถบคอลัมน์ด้านซ้าย คลิกเลือก CRAN



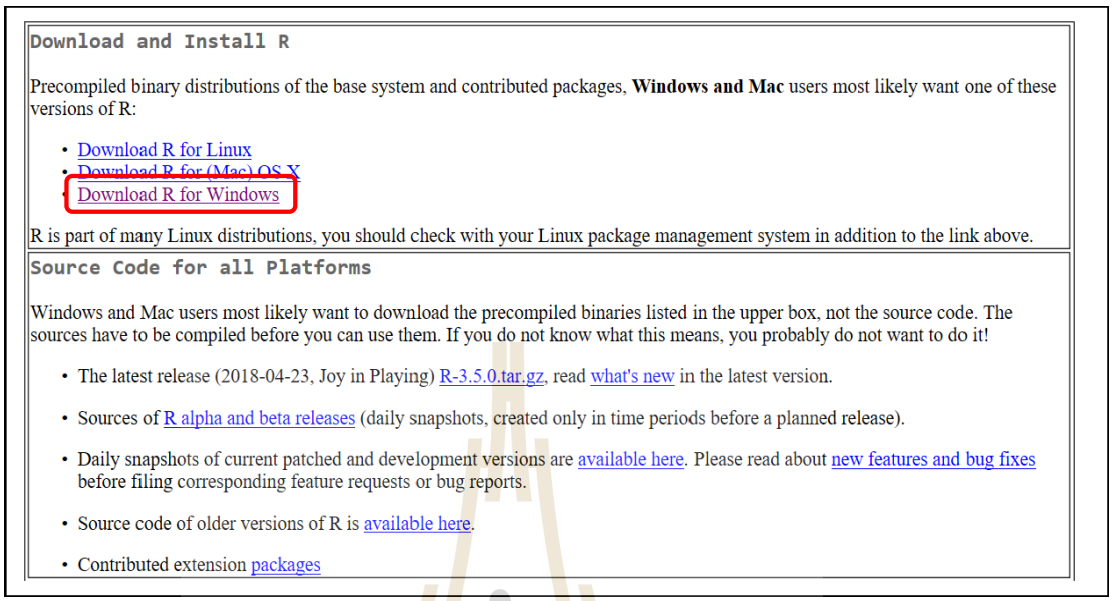
รูปที่ ก.1 หน้าต่างของเว็บไซต์

2) มีแถบรายชื่อประเทศให้เลือก คลิกเลือก Thailand



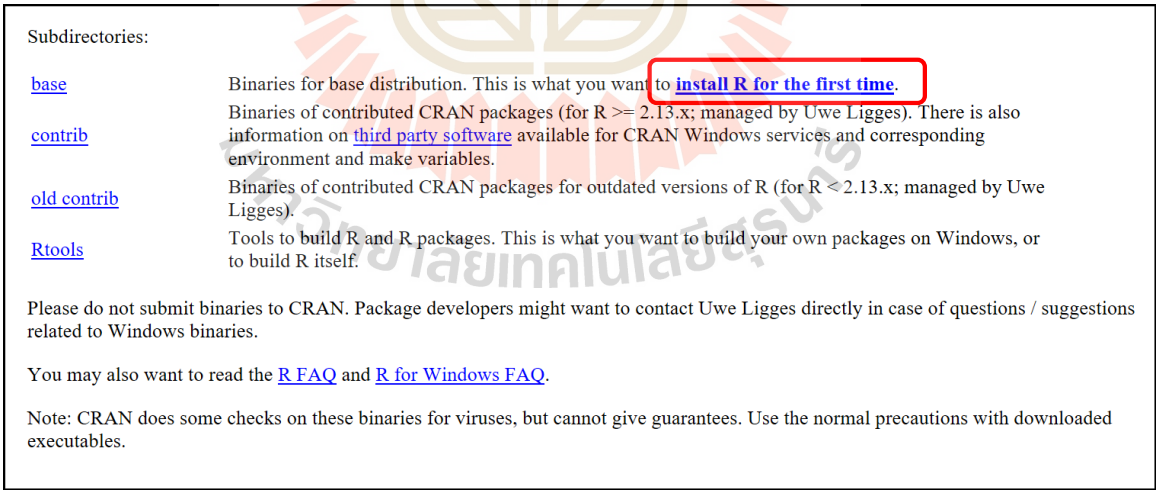
รูปที่ ก.2 แถบแสดงเพื่อเลือกประเทศ

3) คลิก Download สำหรับระบบปฏิบัติการ Windows (Download R for Windows)



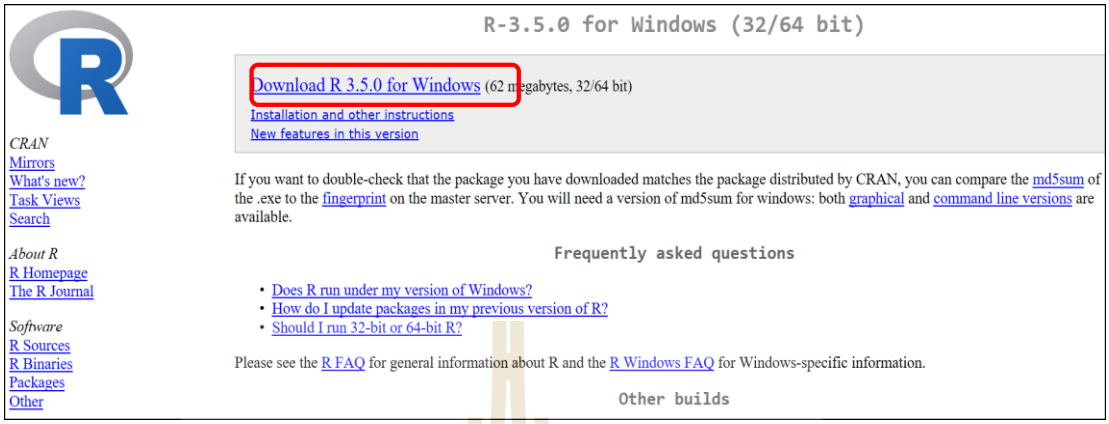
รูปที่ ก.3 แถบดาวน์โหลดโปรแกรม R สำหรับระบบปฏิบัติการ Windows

4) คลิก install R for the first time. ในหัวข้อ base สำหรับผู้ติดตั้งโปรแกรม R ครั้งแรก



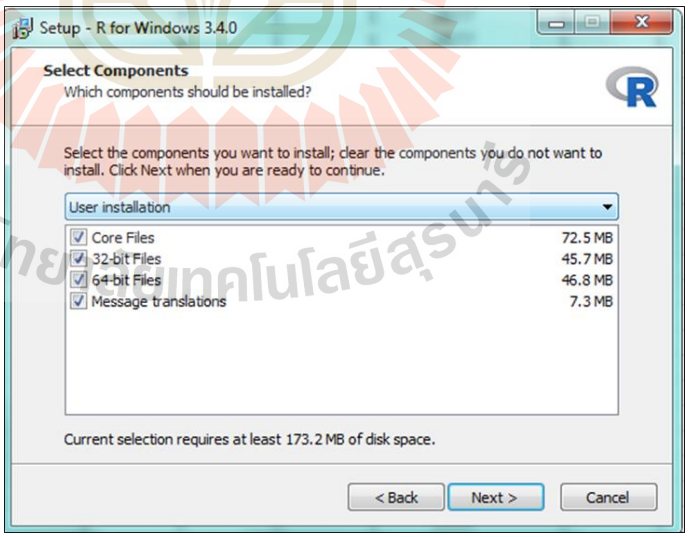
รูปที่ ก.4 แถบการติดตั้งโปรแกรม

5) คลิก Download R 3.4.0 for Windows



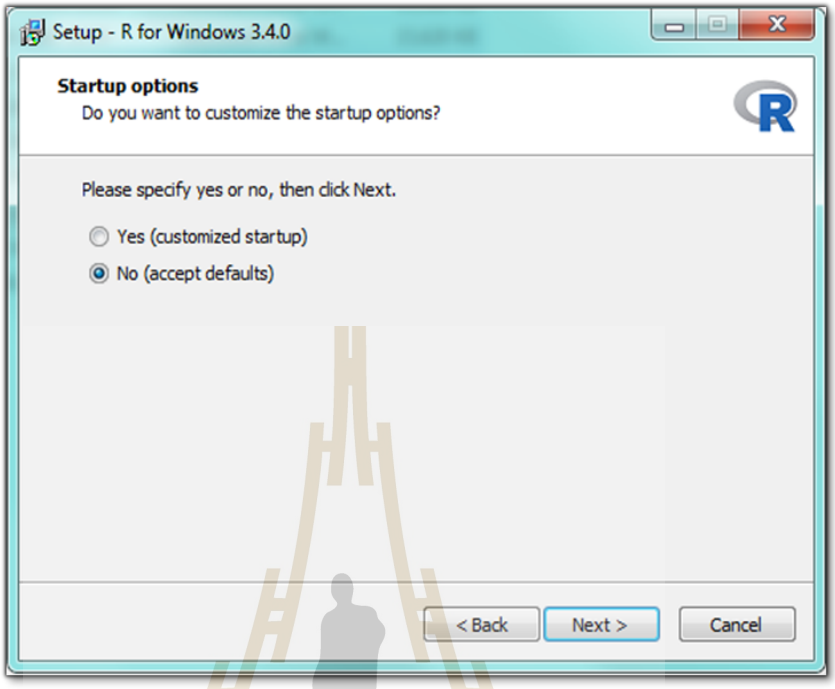
รูปที่ ก.5 แถบดาวน์โหลดโปรแกรม R

6) เมื่อ Download เสร็จสมบูรณ์ ทำการติดตั้ง โดยการ RUN โปรแกรม โดยโปรแกรม R จะเลือกระบบที่เหมาะสมกับคอมพิวเตอร์ แต่เพื่อความมั่นใจ สามารถเลือกได้ว่า จะใช้ระบบ 32 bit หรือ 64 bit แล้วกด Next



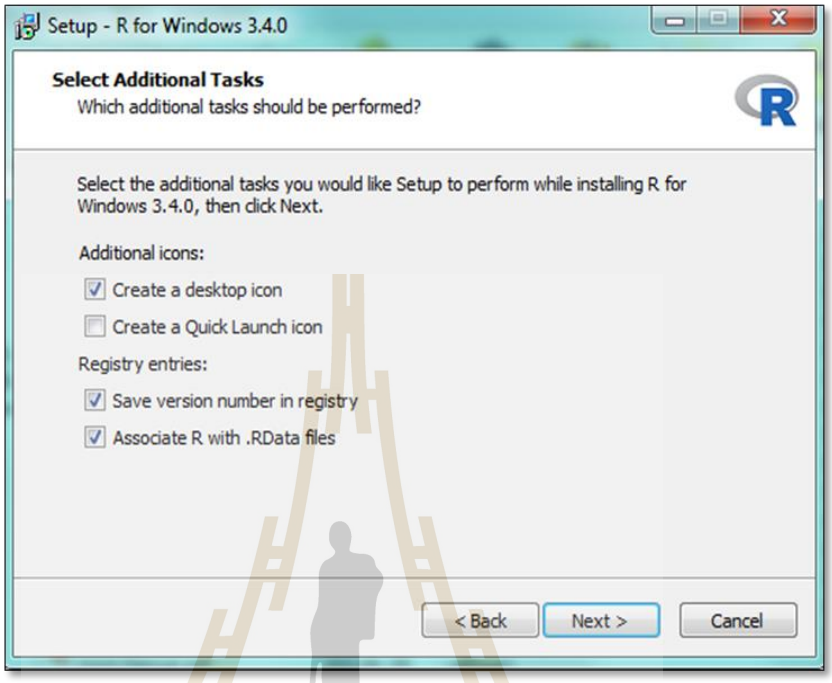
รูปที่ ก.6 หน้าต่างในการติดตั้งโปรแกรม

7) คลิกเลือก No (accept defaults) เพื่อให้ติดตั้งโปรแกรมตามค่าเริ่มต้น แล้วกด Next



รูปที่ ก.6 หน้าต่างในการติดตั้งโปรแกรม (ต่อ)

8) เลือกเครื่องหมายถูกหน้าช่อง Save version number in registry และ Associate R with .RData files ตามค่า defaults ที่โปรแกรมกำหนด และกด Next เพื่อติดตั้งโปรแกรม



รูปที่ ก.6 หน้าต่างในการติดตั้งโปรแกรม (ต่อ)

ภาคผนวก ข

บทความที่ได้รับตีพิมพ์เผยแพร่

รายชื่อบทความที่ได้รับการตีพิมพ์งานวิจัยนี้

Preuksarat, S. and Samattapapong, N. (2018). **Relationships Analysis for Process Variables That Affect Data Storage Device Quality.** International Conference on Mechanical Engineering and Industrial Automation. 23-24 March 2018. Pattaya, Thailand. pp. 156-162.

รายชื่อบทความที่ได้รับการตีพิมพ์ในระหว่างการศึกษา

พฤกษารัตน์ สิทธิพงค์ และ นรา สมัตถภาพงศ์. (2018). การปรับปรุงประสิทธิภาพแถวคอยในการรอรับบริการแผนกเวชระเบียนของศูนย์การแพทย์ มทส. โดยใช้การจำลองสถานการณ์. **วารสารข่ายงานวิศวกรรมอุตสาหการไทย.** 4(1)

4th International Conference On Mechanical Engineering and Industrial Automation Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

Relationships Analysis for Process Variables That Affect Data Storage Device Quality

Preuksarat Sittipong
School of Engineering, Industrial Engineering,
Suranaree University of Technology. S
111 Suranaree University Avenue,
Muang district, Nakhon Ratchasima 30000.

Nara Samattapapong
School of Engineering, Industrial Engineering,
uranaree University of Technology.
111 Suranaree University Avenue
Muang district, Nakhon Ratchasima 30000.

Abstract— The purpose of this research was to study the relationship between variables that affect data storage device quality. There are 39,605 samples of data storage device from 74 variables. There were 53 quantitatives variables and 21 qualitatives variables. The research tools is R Program. Use Pearson Correlation and Mutual Information to find the correlation between the data. The results of this research show that the relationship between the two method given the similarity factor. Comparing Pearson Correlation and Mutual Information. The Mutual Information method can be more accurately described than the Pearson Correlation method.

Keywords— Data Storage Device, Failure, Quality

I. INTRODUCTION

Currently, the data storage device industry is highly competitive. Case Study, a data storage device manufacturing company, wants to optimize the production process. By the loss of the data storage device manufacturing process is caused by some data storage device products not through the process of inspection. Manufacturing process of data storage device, with a complex and massive data, thus making it difficult to verify quality. Data storage device manufacturing process in the quality inspection process is measured by various processes in the production process. These variables affect the quality of the data storage device to pass the quality inspection process. Variables in the manufacturing process, which includes a variety of quantitative variables and qualitative variable. If the manufacturer knows the main parameters that affect the quality of the data storage device in each production cycle. Can analyze the problem and find out why the data storage device failed to pass the test. This will lead to improved quality of the production process and will increase the volume of tested data storage device.

A. Related Theory

1) Correlation: Correlation is a study of relationships between two or more variables (or two sets of data). The degree or magnitude of the relationship will be used the number of correlation coefficients. If the correlation coefficient is near -1 or 1 indicates a high degree of correlation. But the value close to 0 indicates a low or no correlation.

+, - The numbers page correlates to the direction of the relationship. If so

r is marked with + means the relationship is in the same direction. (One variable is high The other one is high.)

r marked - means to have a relationship in the opposite direction (One variable is high Another variable is low.)

4th International Conference On Mechanical Engineering and Industrial Automation Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

Except for some correlation coefficients that have the characteristic $0 \leq r \leq 1$, only the size or level of the relationship is known. Can not tell the direction of the relationship.

Pearson's Correlation Coefficient

Pearson's Correlation Coefficient use the symbol r_{xy} A method for measuring the relationship between two variables or data. The two variables or data must be in the form of data in the Interval or Ratio scale. The formula is calculated as follows.

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

When r_{xy} is Pearson's Correlation Coefficient.

$\sum X$ is sum of data measured from variable 1 (X).

$\sum Y$ is sum of data measured from variable 2 (Y).

$\sum XY$ is sum of multipliers between variables 1 and 2.

$\sum X^2$ is sum of the squares of data measured from variable 1.

$\sum Y^2$ is sum of the squares of data measured from variable 2.

N is size of the sample.

Pearson's Correlation Coefficient use only suitable for data that has a linear relationship. Therefore, in the calculation, if r = 0 interpretation of unrelated data may not be accurate. Since it is possible that the data is related in a non-linear manner.

2) Mutual Information: The mutual information method does not require the development of a forecasting model before use and can be considered a model-free method (unlike a model-based method, which requires the model to be developed beforehand). The model-free method is simpler because it requires less time to screen the data because there is no need to run a model. MI has been developed based on the principles of information theory and the notion of entropy proposed by Shannon (1948) in [1]. The mutual information equation for bivariate data is shown below.

$$MI = \frac{1}{N}\sum_{t=1}^{N} \ln\left[\frac{P_{x,y}(x_i,y_i)}{P_x(x_i)P_y(y_i)}\right] \qquad (1)$$

Where x_i and y_i are the bivariate sample pair.

N is the sample size.

$P_{x,y}(x_i, y_i)$ is the join probability density at the sample point.

$P_x(x_i)$ and $P_y(y_i)$ are the univariate marginal probability densitiesat the sample point, respectively.

Gray R.M (1990) If the value of MI is 0, then no correlation. But If the value of MI is the highest, the data is most likely to be correlated in [2].

3) R Program: R Program is open source software that is free to use, can be adapted as needed. The beginners of the program are Robert Gentleman and Ross Ihaka. R Program is a free program available on various computers. Whether used on Windows, Mac OS, or Linux. The R language developed from the S language developed for use in statistical applications. The R program is an interesting alternative. Because R is an open source, high-performance data analysis and a lot of packages to choose from. The program is copyrighted but free of charge. Can be downloaded anywhere in the world.

B. Related Research

From the study of related research on data storage device quality forecasting to increase productivity, there are various methods of discovery. Each research will have a different way as follows:

4th International Conference On Mechanical Engineering and Industrial Automation Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

Nara Samattapapong (2559) The system of forecasting data storage device production using the GRNN method, there are three additional methods includes Particle Swarm Optimization (PSO), Unrestricted Search Optimization (USO) and Interval Halving Optimization (IHO). To bring the results of all the comparative methods. Forecasting the system using GRNN is the best way to provide predictable results close to the actual value and the next best way is the IHO method in [3].

Joseph F. Murray and group (2005) Comparison of machine learning methods using data from internal attributes of each drive. There are 3 ways to compare includes support vector machines (SVMs), unsupervised clustering and non-parametric statistical tests (rank-sum and reverse arrangements). The best way is non-parametric statistical tests (rank-sum and reverse arrangements) in [4]. Yu Wang and group (2014) Using a two-step parametric method in forecasting failure hard drives. Based on the findings, failures can be detected as 68% and the error rate is 0% this results in better support vector machine and hidden Markov models in [5]. Thanadon Suchatpong and Krischonme Bhumkittipich (2014) Predict the production of failure data storage device using Decision Tree Learning and data collection process in [6]. In the study of research related to the use of the R program to find the relationship between the variables is PhD Candidate Ștefan Cristian CIUCU (2014) A statistical analysis was performed using R and PHP. Case study of foreign direct investment and GDP of the Republic of Moldova in [7]. This research does not predict data storage device quality. However, it is studying the variables that affect the quality of the data storage device. Researcher did not find the research on finding the variables that affect the quality of the data storage device using the R program.

II. RESEARCH METHODOLOGY

This study investigates the variables that affect the quality of the data storage device of a company. The information obtained from the case study company can not be disseminated. Only the methods and results of the research are shown. The steps of the research are shown in Fig. 1.

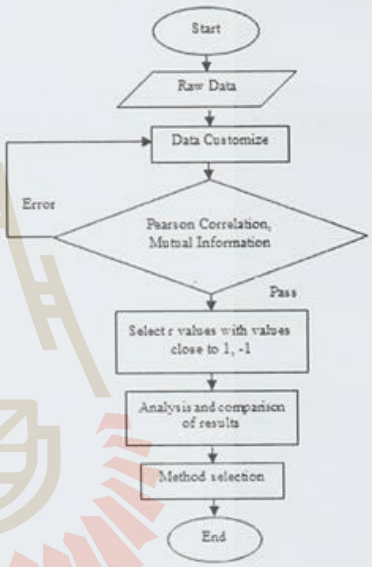


Fig. 1 Shown Research Methodology

Fig. 1 shown the research process from the beginning. Once the information from the company. Researchers have

4th International Conference On Mechanical Engineering and Industrial Automation Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

adapted the information. Cut the results of the data storage device quality testing process with 74 incomplete variants. Only 74 results are available for the next process is R Program. Finding r values are Pearson Correlation and Mutual Information, which show the procedure below.

A. Step to find the r value by Pearson Correlation Method.

1) Open the R program.

2) Import data from Excel program into R program. Go to : Data > Import data > From excel file select OK. The screen show pop up, select OK. Then select the desired excel file and press the Open button. (Available in .xlsx and .csv)

3) Validate by select View data set.

4) Calculate the relationship of data by Pearson method. Go to : Statistics > Summaries > Correlation matrix.

5) Select all the data to find correlation, and select Pearson, then select OK .

6) The program calculates the correlation (r values) to all Matrix in the Output columns.

7) End of calculation.

B. Step to find the r value by Mutual Information Method.

1) Open the R program, the program will show the Packages Info Theo window. Enter the command in the R Console.

2) Import data from excel to R program. Need to save the excel file into the .csv format. (Because this method imports .csv format only.) Type the following command: library(readr) data <- read_csv("Address of file / file name excel.csv")

3) Enter the command to calculate the MI : mutinformation(data$column name attributes, data$OUTPUT, method="emp")

4) Enter the original command, but change the column name until the attribute column is full. Change column data$A_01 to data$A_02 until all 74 variables.

5) End of calculation.

III. RESULTS

A. Result of Pearson Correlation method.

TABLE I

Shown the r value by Pearson method.

Attributes	r Value	Attributes	r Value	Attributes	r Value
A_01	0.029297 73547	A_26	0.01254 03228	C_51	0.0183806288
A_02	0.002771 6949	A_27	0.00037 32049	A_52	0.0012387501
A_03	0	A_28	1.0000000000	C_53	0.0196946031
A_04	0.0159921691	A_29	0	A_54	0.0215349134
A_05	0.0156894409	A_30	0	C_55	0.0054760183
A_06	0.0249775981	A_31	0	C_56	0.0245853087
A_07	1.0000000000	A_32	0	C_57	0.0006875783
A_08	0	C_33	0.2154517521	C_58	0.0060900156
A_09	0	C_34	0.00 3229028	C_59	0.0049528754
A_10	0	C_35	0.0092256399	C_60	0.0075083970
A_11	0.00029 56563	C_36	0.01557 4740	A_61	0
A_12	0.03567 65028	C_37	0.3308613004	A_62	0.0257163929
A_13	0.0256067331	C_38	0.2358155305	A_63	0.0256067331
A_14	0	A_39	0.0154125837	A_64	0.0256067331
A_15	0	A_40	0.0253481595	A_65	0.0204688057
A_16	0.0054 25844	A_41	0	C_66	0.0025311672
A_17	0.0256067331	A_42	0.0241 68866	C_67	0.0134958088
A_18	0.9506 72553	A_43	0.0227 89496	C_68	0.0239906871
A_19	1.0000000000	A_44	0	A_69	0.0062 76020
A_20	1.0000000000	A_45	0	A_70	0.0253567 29
A_21	1.0000000000	A_46	0	A_71	0
A_22	0.0012 58544	C_47	0.0036317039	A_72	0.0117 95504
A_23	0.0028 30 82	C_48	0.0102 95065	A_73	0.0068 95440
A_24	0.0139 58653	C_49	0.0053 14429	A_74	0.0028 56587
A_25	0.0028 46468	C_50	0.0022 86535		

Calculated by Pearson Correlation. If the relation of the data is close to 1 or -1, then the data is very closely related.

4th International Conference On Mechanical Engineering and Industrial Automation Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

From calculate the variables that affect the quality of the work. Can sort the relationship as shown in Fig. 2.

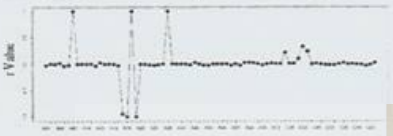


Fig. 2 Shown relationship between variables with value r by Pearson correlation method.

B. Result of Mutual Information Method

Table II

Shown calculation MI value by Mutual Information

Attribute	MI Value	Attribute	MI Value	Attribute	MI Valu
A_01	0.009519488	A_26	0.00304748	C_51	Error
A_02	3.22E-05	A_27	0.000780168	A_52	Error
A_03	0	A_28	0.009519488	C_53	Error
A_04	0.000845475	A_29	0	A_54	Error
A_05	0.008131310	A_30	0	C_55	Error
A_06	0.000526011	A_31	0	C_56	Error
A_07	0.009519488	A_32	0	C_57	2.52E-0
A_08	0	C_33	Error	C_58	Error
A_09	0	C_34	Error	C_59	Error
A_10	0	C_35	Error	C_60	Error
A_11	4.51E-08	C_36	0.008124255	A_61	0
A_12	0.001095786	C_37	0.006514698	A_62	0.0003511
A_13	0.000526011	C_38	Error	A_63	0.0005260
A_14	0	A_39	0.003597721	A_64	0.0005260
A_15	0	A_40	0.00652668	A_65	0.0005250
A_16	0.009519488	A_41	0	C_66	Error
A_17	0.000526011	A_42	0.000938156	C_67	0.0012499
A_18	0.009519488	A_43	0.003444311	C_68	0.0005260
A_19	0.009594880	A_44	0	A_69	3.51E-0
A_20	0.009519488	A_45	0	A_70	0.0008434
A_21	0.009519488	A_46	0	A_71	0
A_22	0.009519488	C_47	Error	A_72	0.0001836
A_23	0.000645846	C_48	Error	A_73	2.13E-0
A_24	0.001328113	C_49	Error	A_74	0.0009771
A_25	0.000892669	C_50	Error		

Calculated by Mutual Information if the value of MI is 0, then no correlation. But If the value of MI is the highest, the data is most likely to be correlated. The error is caused by the data type is discontinuous as shown in Fig. 3.

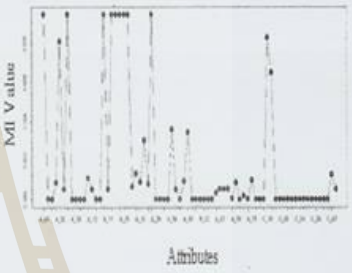


Fig. 3 Shown relationship between variables with MI value by Mutual information.

IV. CONCLUSION

Analysis of the relationship between variables affecting the quality of data storage device drive production using Pearson Correlation and Mutual Information. It was found that the correlation coefficient and data storage device quality in each method gave similar correlation coefficient.

4th International conference On Mechanical Engineering And industrial Automation
Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

The variables that affect the quality of the data storage device are the highest value of MI, which has 10 attributes, is shown in Table IV.

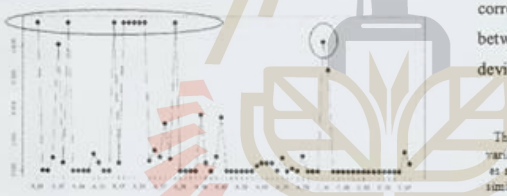
Table IV

Variables that affect the quality of data storage device by Mutual Information

No.	Attribute	MI Value
1	A_01	0.009519488
2	A_07	0.009519488
3	A_16	0.009519488
4	A_18	0.009519488
5	A_19	0.009519488
6	A_20	0.009519488
7	A_21	0.009519488
8	A_22	0.009519488
9	A_28	0.009519488
10	C_36	0.008324255

No.	Attribute	r Value
1	A_07	1.0000000000
2	A_18	-0.9506825353
3	A_19	-1.0000000000
4	A_20	1.0000000000
5	A_21	-1.0000000000
6	A_28	1.0000000000
7	C_33	0.2134517521
8	C_36	0.1015574740
9	C_37	0.3303615004
10	C_38	0.2358155205

B. Results of Operations by Mutual Information



The variables that affect the quality of the data storage device are the r values closest to 1, -1. There are 10 attributes shown in Table III.

Table III

Variables that affect the quality of data storage device by Pearson Correlation.

Summary of Operations between Pearson Correlation and Mutual Information. The Mutual Information can be more accurately described than the Pearson Correlation. The correlation between the two methods gives the correlation between the variables and the quality of the data storage device is the same 7 variables as the table V.

Table V

The variables are similar

A_07	A_18	A_19	A_20	A_21	A_28	C_36

in both methods

V. SUGGESTION

A. Finding correlation values for both methods yielded similar results. Therefore, choose the appropriate method of data.

B. Mutual Information can find more precise relationships of data, but data must be discrete.

4th International conference On Mechanical Engineering And industrial Automation
Held on 23rd-24th March 2018, in Pattaya, Thailand ISBN: 9780998900049

VI. REFERENCES

[1] Shannon, *A mathematical theory of communication*. Bell System Technical Journal, pp. 379-423, 1948.

[2] Gray R.M., *Entropy and information theory*, 284p, 1990.

[3] Nara Samattapapong, *A production Throughput Forecasting System in an Automated Hard Disk Drive Test Operation ing GRNN*, Journal of Industrial Engineering and Management, vol. 9, pp. 330-358, 2559.

[4] Joseph F. Murray, Gordon F. Hughes and Kenneth Kreutz-Delgano, *Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application*, Journal of Machine Learning Research, pp. 783–816, 2005.

[5] Yu Wang, Eden W. M. Ma, Tommy W. S. Chow and Kwok-Leung Tsui, A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives, *IEEE Transactions on Industrial Informatics*., vol. 10, pp. 430-419, Feb. .2014

[6] PhD Candidate Ştefan Cristian CIUCU, *Statistical Data Analysis via R and PHP: A Case Study Of the Relationship Between GDP and Foreign Direct Investments for The Republic Of Moldova*, Romanian Statistical Review, 2014.

[7] Thanadon Suchatpong and Krischonme Bhumkittipich, *Hard Disk Drive Failure Mode Prediction based on Industrial Standard using Decision Tree Learning*, IEEE 978-1-4799-2993-1, 2014.

ประวัติผู้เขียน

นางสาวพฤกษารัตน์ สิทธิพงค์ เกิดเมื่อวันที่ 7 พฤษภาคม พ.ศ. 2537 จบการศึกษาระดับชั้นประถมศึกษาปีที่ 1 จนถึงมัธยมศึกษาปีที่ 3 ที่โรงเรียนเทพวิทยา ตำบลกรับใหญ่ อำเภอบ้านโป่ง จังหวัดราชบุรี ระดับชั้นมัธยมศึกษาปีที่ 4-6 ที่โรงเรียนประสาทรัฐประชากิจ ตำบลประสาทสิทธิ์ อำเภอดำเนินสะดวก จังหวัดราชบุรี สำเร็จการศึกษาระดับปริญญาตรี สำนักวิชาวิศวกรรมศาสตร์ สาขาวิศวกรรมอุตสาหการ มหาวิทยาลัยเทคโนโลยีสุรนารี ในปีการศึกษา 2558 ต่อมาได้เข้าศึกษาต่อในระดับปริญญาโท สำนักวิชาวิศวกรรมศาสตร์ สาขาวิศวกรรมระบบอุตสาหกรรมและสิ่งแวดล้อม ในปีการศึกษา 2559