บทที่ 1

บทนำ

1.1 คำนำ

การวิเคราะห์ตัวแปรพหุ (MULTIVARIATE ANALYSIS) ซึ่งเป็นการวิเคราะห์ข้อมูลที่มีตัวแปรสุ่ม (RANDOM VARIABLE) หลาย ๆ ตัว นิยมใช้กันมากทางด้านจิตวิทยา การศึกษา การแพทย์ สังคมศาสตร์ และด้านอื่น ๆ โดยเฉพาะอย่างยิ่งในการหาความสัมพันธ์ระหว่างตัวแปร การวัดความสัมพันธ์ดังกล่าวมีวิธีวิเคราะห์ได้ เช่น การวิเคราะห์จำแนกประเภท (DISCRIMINANT ANALYSIS) การวิเคราะห์ความสัมพันธ์แคนอนิคัล (CANONICAL CORRELATION) และการวิเคราะห์ตัวประกอบ (FACTOR ANALYSIS) เป็นต้น ในการวิเคราะห์ตัวแปรพหุมีข้อตกลงเบื้องต้นที่สำคัญคือ การแจกแจงร่วม (JOINT DISTRIBUTION) ของตัวแปรสุ่มต้องมีการแจกแจงเป็นแบบปกติ และค่าสังเกตแต่ละค่าต้องเป็นอิสระต่อกัน

ในการวิเคราะห์ตัวแปรพหุไม่ว่าวิธีวิเคราะห์ใด ๆ ก็ตาม ถ้าข้อมูลที่รวบรวมมาเพื่อทำการวิเคราะห์เป็นไปตามข้อตกลงเบื้องต้น และมีข้อมูลครบสมบูรณ์ทุกตัวก็คงไม่เกิดปัญหาในการวิเคราะห์นั้น แต่ถ้าหากข้อมูลที่รวบรวมมาได้นั้นมีข้อมูลบางตัวสูญหายไป และไม่สามารถตามไปเก็บเพิ่มเติมได้ จึงทำให้ข้อมูลของตัวอย่างบางตัวอย่างไม่สมบูรณ์ ทำการวิเคราะห์ไม่ได้ นอกจากว่าผู้วิจัยจะตัดตัวอย่างชุดนั้นทิ้งไป ซึ่งจะมีผลทำให้ขนาดตัวอย่างมีจำนวนลดน้อยลง และที่สำคัญยิ่งก็คือทำให้สูญเสียรายละเอียดของข้อมูลบางชุดไป ซึ่งอาจจะมีผลกระทบต่อผลสรุปของการวิเคราะห์นั้น ๆ ได้ เพื่อหลีกเลี่ยงการสูญเสียรายละเอียดของข้อมูลบางชุดไป ควรจะมีการประมาณค่าสังเกตที่สูญหายไปเสียก่อน แล้วจึงนำค่าสังเกตเหล่านั้นไปทำการวิเคราะห์ ดังนั้นวิธีการประมาณค่าสังเกตที่สูญหายไปจึงมีความสำคัญมาก แต่เนื่องจากวิธีการประมาณค่าสังเกตที่สูญหายไปนั้นมีหลายวิธี วิทยานิพนธ์นี้จึงสนใจที่จะทำการเปรียบเทียบวิธีการประมาณค่าสังเกตที่สูญหายไป ว่าวิธีการใดจะเป็นวิธีประมาณที่ดีที่สุด เมื่อข้อมูลมีลักษณะต่าง ๆ กัน

1.2 ที่มาของปัญหา

เนื่องจากปัญหาการสูญหายของข้อมูล อาจมีผลกระทบต่อผลสรุปของการวิเคราะห์ได้ ถ้าหากต้องตัดตัวอย่างนั้น ๆ ทิ้งไป จึงได้มีนักสถิติหลายท่านพยายามคิดค้นวิธีการประมาณค่าสูญหายของการวิเคราะห์แบบต่าง ๆ ขึ้น กล่าวคือ

ค.ศ. 1966 - ค.ศ. 1969 A.A AFIFI และ R.M. ELASHOFF ได้ศึกษาเปรียบเทียบ วิธีการประมาณค่าสูญหายจากการวิเคราะห์ความถดถอยพหุเชิงเส้น (MULTIPLE LINEAR REGRESSION) โดยใช้ค่าเฉลี่ยความคลาดเคลื่อนเป็นเกณฑ์ในการเปรียบเทียบ

ค.ศ. 1968 E.C. JACKSON ได้ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์จำแนกประเภท (DISCRIMINANT ANALYSIS) โดยใช้ข้อมูลจากตัวอย่าง 1 ชุด และใช้ ร้อยละของการจำแนกผิด เป็นเกณฑ์ในการเปรียบเทียบ

ต่อมา ค.ศ. 1972 LINDA S. CHAN และ OLIVE JEAN DUNN ได้ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์จำแนกประเภทในกรณี 2 ประชากรทั้งหมด 5 วิธี คือ

ก. ศึกษาเมื่อไม่มีข้อมูลสูญหายเลย

ข. ศึกษาเมื่อมีข้อมูลสูญหายโดยตัดตัวอย่างที่มีข้อมูลสูญหายออก

ค. ศึกษาเมื่อมีข้อมูลสูญหายโดยประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (MEAN)

ง. ศึกษาเมื่อมีข้อมูลสูญหายโดยประมาณข้อมูลสูญหายด้วยวิธีวิเคราะห์ความถดถอยพหุเชิงเส้น (MULTIPLE LINEAR REGRESSION)

จ. ศึกษาเมื่อมีข้อมูลสูญหายโดยประมาณข้อมูลสูญหายด้วยวิธีวิเคราะห์ส่วนประกอบหลัก (PRINCIPAL COMPONENT)

โดยใช้เทคนิค MONTE CARLO SIMULATION ภายใต้ข้อตกลงเบื้องต้นว่า

ก. ทั้ง 2 ประชากรมีความแปรปรวนร่วมเท่ากัน

ข. ตัวแปรสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ

ค. ค่าสังเกตแต่ละค่าเป็นอิสระจากกัน

ง. การสูญหายเกิดขึ้นโดยสุ่มในแต่ละตัวแปรและในแต่ละตัวอย่าง

จ. เปอร์เซ็นต์ของการสูญหายของแต่ละตัวแปรใกล้เคียงกัน

และใช้ร้อยละของการจำแนกผิด เป็นเกณฑ์ในการเปรียบเทียบวิธีต่าง ๆ ข้างต้น

ต่อมา ค.ศ. 1974 LINDA S. CHAN และ OLIVE JEAN DUNN ได้ศึกษาเพิ่มเติมในกรณีที่ขนาดตัวอย่างใหญ่ ภายใต้ข้อสมมุติเดิม

ค.ศ. 1976 LINDA S. CHAN, JUNE AONO GILMAN และ OLIVE JEAN DUNN ได้ศึกษาเพิ่มเติมอีกโดยเพิ่มวิธีการประมาณอีก 2 วิธีคือ วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นดัดแปลง (MODIFIED MULTIPLE LINEAR REGRESSION) และวิธีการวิเคราะห์ส่วนประกอบหลักดัดแปลง (MODIFIED PRINCIPAL COMPONENT) ภายใต้ข้อสมมุติเดิม

ค.ศ. 1978 RODERICK J.A. LITTLE ได้ศึกษาเพิ่มเติมโดยปรับปรุงวิธีการวิเคราะห์ความถดถอยพหุเชิงเส้น และวิธีการวิเคราะห์การถดถอยพหุเชิงเส้นดัดแปลง แต่เป็นวิธีการที่ค่อนข้างยุ่งยากและใช้เวลานาน

จะเห็นว่าจากการวิจัยที่ผ่านมาสามารถนำผลสรุปไปใช้ประมาณค่าสูญหายในการวิเคราะห์ความถดถอยพหุเชิงเส้นและการวิเคราะห์จำแนกประเภทเท่านั้น แต่ในทางปฏิบัติ การวิเคราะห์ข้อมูลหลายตัวแปรมีการวิเคราะห์หลายวิธี ดังนั้นถ้าหากเกิดปัญหาข้อมูลสูญหายก็ยังไม่สามารถนำผลสรุปเหล่านั้นไปใช้ได้ ผู้วิจัยจึงมีความสนใจจะศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ โดยศึกษาในกรณีที่ข้อมูลเป็นไปตามข้อตกลงเบื้องต้น เพื่อประโยชน์สำหรับผู้วิจัยอื่น ๆ ที่ศึกษาเกี่ยวกับการวิเคราะห์ตัวแปรพหุ และมีปัญหาข้อมูลสูญหายเกิดขึ้น จะสามารถเลือกวิธีการประมาณค่าสูญหายได้สอดคล้องกับลักษณะข้อมูลที่มีอยู่

วิธีประมาณค่าสูญหายที่สนใจนำมาเปรียบเทียบในการทำวิทยานิพนธ์นี้คือ

1. วิธีที่ใช้ค่าเฉลี่ย
2. วิธีวิเคราะห์ความถดถอยพหุเชิงเส้น
3. วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นดัดแปลง
4. วิธีวิเคราะห์ส่วนประกอบหลัก

1.3 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ
2. เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุทั้ง 4 วิธี

โดยพิจารณาจากค่าเฉลี่ยความคลาดเคลื่อน (MEAN SQUARE ERROR) และคะแนนรวมจากการถ่วงน้ำหนักด้วย 4, 3, 2, 1 ของจำนวนครั้งของวิธีทั้ง 4 ที่ได้ลำดับที่ 1, 2, 3, 4 ตามลำดับ เมื่อเรียงลำดับค่า MSE จากน้อยไปหามาก

1.4 ขอบเขตของการวิจัย

1. ในการศึกษาเปรียบเทียบนี้จะเปรียบเทียบวิธีการประมาณค่าสูญหาย 4 วิธีคือ วิธีที่ใช้ค่าเฉลี่ย วิธีวิเคราะห์ความถดถอยพหุเชิงเส้น วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นตัดแปลง และวิธีวิเคราะห์ส่วนประกอบหลัก

2. ประชากรที่นำมาศึกษามีการแจกแจงแบบปกติ (MULTIVARIATE NORMAL DISTRIBUTION) $\sim N(\underset{\sim}{\mu}, \Sigma)$ โดยที่

$$\underset{\sim}{\mu} = \underset{\sim}{0}$$

$$\Sigma = \rho_{ij} \quad , \quad \rho_{ij} = \begin{cases} \rho & \text{เมื่อ } i \neq j \\ 1 & \text{เมื่อ } i = j \end{cases}$$

เมื่อ ρ_{ij} คือความสัมพันธ์ระหว่างตัวแปรที่ i และตัวแปรที่ j และ

ρ มีค่า 3 ระดับคือ 0.2 0.5 0.8

3. จำนวนตัวแปรที่สนใจศึกษามี 4 ระดับคือ $p = 3$, $p = 5$, $p = 7$ และ $p = 10$

4. ขนาดตัวอย่างที่ใช้มี 5 ระดับคือ $n = 30$ $n = 50$ $n = 70$ $n = 100$ $n = 200$

5. การสูญหายของข้อมูลเป็นไปโดยสุ่ม

6. สัดส่วนของข้อมูลสูญหายของแต่ละตัวแปรมีค่าใกล้เคียงกัน คือ 10%

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ทำให้ทราบวิธีที่ดีที่สุด สำหรับการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ เมื่อจำนวนตัวแปร ขนาดตัวอย่าง และความสัมพันธ์ระหว่างตัวแปรที่ใช้ในการวิเคราะห์แตกต่างกัน