รายงานวิจัยฉบับสมบูรณ์

การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

Development of Simple Linear Regression Models by Using Balance Ranked Set Sampling

โดย

ชฎารัตน์ ถาปัน (หัวหน้าโครงการ) คณะวิทยาศาสตร์และเทคโนโลยี

สลิลทิพย์ แดงกองโค (ผู้ร่วมวิจัย) คณะวิทยาศาสตร์และเทคโนโลยี

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยราชภัฏพิบูลสงคราม

2562

แบบสรุปโครงการวิจัยฉบับสมบูรณ์

---

1. ชื่อโครงการ (ภาษาไทย) การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

(ภาษาอังกฤษ) Development of Simple Linear Regression Models by Using Balance Ranked Set Sampling.

2. ชื่อหัวหน้าโครงการ นางสาวชฎารัตน์ ถาปัน

Miss Chadarat Tapan

ตำแหน่ง อาจารย์

หน่วยงาน คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏพิบูลสงคราม

สถานที่ติดต่อ สำนักงานคณบดีคณะวิทยาศาสตร์และเทคโนโลยี 156 หมู่ 5 ต.พลายชุมพล อ.เมือง จ.พิษณุโลก 65000

โทรศัพท์ 0897571729 E-mail N_nstat@hotmail.com

ผู้ร่วมโครงการ นางสาวสลิลทิพย์ แดงกองโค

Miss Salinthip Daengkongkho

ตำแหน่ง อาจารย์

หน่วยงาน คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏพิบูลสงคราม

สถานที่ติดต่อ สำนักงานคณบดีคณะวิทยาศาสตร์และเทคโนโลยี 156 หมู่ 5 ต.พลายชุมพล อ.เมือง จ.พิษณุโลก 65000

โทรศัพท์ 0872103803 E-mail dsalinthip@psru.ac.th

3. ระยะเวลาโครงการ 1 ปี ตั้งแต่เดือน กุมภาพันธ์ พ.ศ. 2561 ถึงเดือน กุมภาพันธ์ พ.ศ. 2562

4. บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling: BRSS) และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งได้ดำเนินการจำลองข้อมูลประชากรตามแผนการจำลองข้อมูล จากนั้นทำการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อสร้างตัวแบบถดถอย ซึ่งเกณฑ์ที่ใช้ในการพิจารณาตัวแบบถดถอยคือ ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2) และค่าความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) ผลการศึกษาโดยสรุปพบว่า ตัวแบบถดถอยที่สร้างจากตัวอย่างที่ได้มาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เมื่อมีขนาดตัวอย่างเพิ่มมากขึ้นสำหรับการสุ่มตัวอย่าง 1 รอบ ($m=1$) ตัวแบบถดถอยมีค่าสัมประสิทธิ์การตัดสินใจสูงที่สุด และมีค่าความคลาดเคลื่อนกำลังสองของตัวแบบถดถอยน้อยที่สุด

คำสำคัญ: การสุ่มตัวอย่าง ตัวแบบถดถอยเชิงเส้น การสุ่มตัวอย่างแบบชุดลำดับ

Abstract

The purposes of this study are to propose simple linear regression model based on balance ranked set sampling and to check the accuracy and reliability of simple linear regression model with balance ranked set sampling. The investigation was done mainly via simulation study. Finite population was generated by simulation plans. Then, ranked set sampling was applied so as to obtain sample data to create a model. The criteria use for consideration of simple linear regression model are coefficient of determination (R^2) and mean square error (MSE). It was found that the simple linear regression model was sampling by balance ranked set sampling when increasing sample size for cycle of sampling at 1 cycle ($m=1$) to obtain a most coefficient of determination and at least MSE.

Keywords: sampling, linear regression model, ranked set sampling

5. วัตถุประสงค์ของโครงการ

1) เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

2) เพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล

6. เป้าหมายของโครงการ

ได้รับผลงานวิจัยเรื่องการสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับตัวแบบถดถอยเชิงเส้นอย่างง่าย

7. งบประมาณ

รายละเอียดงบประมาณ	งบประมาณที่ได้รับ (บาท)	งบประมาณที่ใช้จ่าย (บาท)	ยอดคงเหลือ (บาท)
1. งบบุคลากร	-	-	-
2. งบดำเนินงาน	40,000	40,000	-
2.1 ค่าตอบแทน	-	-	-
2.2 ค่าใช้สอย	16,680	10,660	
-ค่าอาหารและเครื่องดื่ม	2,880	1,440 (240x2คนx3วัน)	
-ค่าเดินทางโดยพาหนะส่วนตัว	7,000	3,620 (362 ก.ม.x5บาทx2 ไปกลับ)	
-ค่าที่พัก	6,800	3,600 (900x2คนx2วัน)	
-ค่าลงทะเบียนงานประชุมวิชาการ	-	2,000	
2.3 ค่าวัสดุ	23,320	29,345	
-วัสดุสำนักงาน	6,000	9,260	
-วัสดุคอมพิวเตอร์	7,000	6,420	
-ค่าหนังสือวารสารและตำรา	4,000	3,697	
-ค่าจัดทำรายงาน (ถ่ายเอกสาร, โปสเตอร์,เข้าเล่ม)	5,000	4,968	
-ค่าจ้างพิมพ์งาน	1,320	5,000	
3. งบลงทุน (ครุภัณฑ์ ถ้ามี)	-	-	-
รวม	40,000	40,005	-

กรณีมีเงินทุนวิจัยคงเหลือให้หัวหน้าโครงการนำเงินคงเหลือพร้อมดอกผล (ถ้ามี) และสมุดบัญชีเงินฝาก (ต้นฉบับ) ส่งหน่วยงานภายใน 30 วัน นับแต่วันสิ้นสุดโครงการ เพื่อให้หน่วยงานตรวจสอบและทำรายงานเสนอมหาวิทยาลัยต่อไป

8. ผลงานที่ได้รับจากโครงการนี้

ให้ผู้วิจัยรายงานผลงานวิจัยตามหัวข้อในตารางรายละเอียดผลงาน ซึ่งประกอบด้วย รูปแบบผลงานวิจัย การผลิตนักศึกษา การจดสิทธิบัตร และการเสนอผลงานวิจัย

ผลงาน	รายละเอียด
1. รูปแบบผลงานวิจัย ได้แก่ ต้นแบบผลิตภัณฑ์/กระบวนการใหม่/เทคโนโลยีใหม่/องค์ความรู้	
❑ ยังไม่ได้รูปแบบผลงานวิจัยที่ชัดเจน	
☑ ได้รูปแบบผลงานวิจัย ดังนี้ (ระบุรายละเอียดโดยย่อของแต่ละรูปแบบ)	
❑ ต้นแบบผลิตภัณฑ์ .... ❑ กระบวนการใหม่ .... ❑ เทคโนโลยีใหม่	1.1 เชิงพาณิชย์ (ระบุชื่อบริษัท/องค์กร/สถาบัน และกิจกรรมโดยย่อในการนำเอาผลงานวิจัยไปใช้) ☐ ก. ดำเนินการแล้ว.. ☐ ข. อยู่ระหว่างดำเนินการ.. ☐ ค. ยังไม่มีการนำผลงานวิจัยไปใช้ประโยชน์เชิงพาณิชย์ ☐ มีแผนที่จะดำเนินการ ในวัน/เดือน/ปี................................ หากต้องการให้มหาวิทยาลัยประสานงานกับภาคเอกชน กรุณาแจ้งให้ ทราบด้วย ☐ อื่น ๆ ระบุ..
.... ☑ องค์ความรู้	1.2 เชิงสาธารณะประโยชน์ (ระบุว่าเป็นกรณีที่ 1 และ/หรือกรณีที่ 2) 1.2.1 <u>กรณีที่ 1</u> เป็นการนำผลงานวิจัยถ่ายทอดให้กับหน่วยงานภาครัฐ/ภาคเอกชน/ชุมชน/กลุ่มบุคคลโดยไม่หวังผลกำไร (ให้ระบุชื่อหน่วยงาน/ชุมชน/กลุ่มบุคคลที่รับผลงานวิจัยไปใช้ประโยชน์และกิจกรรมโดยย่อในการนำผลงานวิจัยไปใช้) 1.2.2 <u>กรณีที่ 2</u> เป็นการเผยแพร่ผลงานวิจัยโดยการจัดประชุม/สัมมนา/ฝึกอบรม (ให้ระบุชื่อหัวข้อที่จัด วัน/เดือน/ปีที่จัด และสถานที่ที่จัด) ❑ ก. ดำเนินการแล้ว.. ❑ ข. อยู่ระหว่างดำเนินการ.. ❑ ค. ยังไม่มีการนำเสนอผลงานวิจัยไปใช้เชิงสาธารณะประโยชน์ ❑ มีแผนที่จะดำเนินการ ในวัน/เดือน/ปี................................ หากต้องการให้มหาวิทยาลัยประสานงานกับภาคเอกชน กรุณาแจ้งให้ ทราบด้วย ❑ อื่น ๆ ระบุ.. หมายเหตุ ถ้ารูปแบบผลงานวิจัยมีมากกว่า 1 รูปแบบให้ระบุการนำไปใช้ประโยชน์ในแต่ละรูปแบบ เช่น โครงการ ก. มี 2 รูปแบบคือ 1) ต้นแบบผลิตภัณฑ์ ให้ระบุการนำไปใช้ประโยชน์ทั้ง 2 ประเภท และ 2) เทคโนโลยีใหม่ ให้ระบุการนำไปใช้ประโยชน์ทั้ง 2 ประเภท ด้วย
2. สิทธิบัตร	
❑ 2.1 จดสิทธิบัตรแล้ว	ระบุรูปแบบผลงานวิจัยที่นำไปจด วัน/เดือน/ปีที่ยื่นจด หมายเลขสิทธิบัตร ประเทศที่ยื่นจดสิทธิบัตร
❑ 2.2 กำลังดำเนินการยื่นขอจดสิทธิบัตร	ระบุรูปแบบผลงานวิจัยที่นำไปจด วัน/เดือน/ปีที่ยื่นจด หมายเลขสิทธิบัตร ประเทศที่ยื่นจดสิทธิบัตร
❑ 2.3 อยู่ในระหว่างเตรียมคำขอจดสิทธิบัตร	ระบุรูปแบบผลงานวิจัยที่นำไปยื่นจด

ผลงาน	รายละเอียด	
☐ 2.4 ยังไม่จดสิทธิบัตร	☐ ก. ต้องการคำปรึกษาจากเจ้าหน้าที่ด้านจดสิทธิบัตรของมหาวิทยาลัย ☐ อื่น ๆ ระบุ..	
3. การเสนอผลงานวิจัย		
☐ 3.1 ยังไม่มีการนำเสนอผลงานวิจัย		
☑ 3.2 มีการนำเสนอผลงานวิจัยแล้วในรูปแบบ ดังนี้		
3.2.1 บทความทางวิชาการ		
☐ 3.2.1.1 วารสาร (Journal)		สถานภาพ
	☐ ก. ระดับชาติ (ระบุชื่อผู้วิจัยและ/หรือผู้ร่วมวิจัย ปีที่ตีพิมพ์ ชื่อบทความ ชื่อวารสาร ฉบับที่ และเลขหน้าที่พิมพ์)	☐ อยู่ระหว่างการเรียบเรียง/เขียน (In preparation) ☐ ยื่นเอกสารแล้ว อยู่ระหว่างการพิจารณา (Submitted) ☐ ได้รับการตอบรับแล้ว อยู่ระหว่างการจัดพิมพ์ (Accepted, In press) ☐ ได้รับการจัดพิมพ์แล้ว (Published)
	☐ ข. ระดับนานาประเทศ (ระบุชื่อผู้วิจัยและ/หรือผู้ร่วมวิจัย ปีที่ตีพิมพ์ ชื่อบทความ ชื่อวารสาร ฉบับที่ และเลขหน้าที่พิมพ์)	☐ อยู่ระหว่างการเรียบเรียง/เขียน (In preparation) ☐ ยื่นเอกสารแล้ว อยู่ระหว่างการพิจารณา (Submitted) ☐ ได้รับการตอบรับแล้ว อยู่ระหว่างการจัดพิมพ์ (Accepted, In press) ☐ ได้รับการจัดพิมพ์แล้ว (Published)
☐ 3.2.1.2 หนังสือ/คู่มือ/ตำรา	☐ ก. ภาษาไทย (ระบุชื่อผู้เขียน ชื่อหนังสือ ชื่อเรื่อง ชื่อสำนักพิมพ์ และวัน/เดือน/ปีที่พิมพ์)	☐ อยู่ระหว่างการเรียบเรียง/เขียน (In preparation) ☐ ได้รับการจัดพิมพ์แล้ว (Published)
	☐ ข. ภาษาอังกฤษ (ระบุชื่อผู้เขียน ชื่อหนังสือ ชื่อเรื่อง ชื่อสำนักพิมพ์ และวัน/เดือน/ปีที่พิมพ์)	☐ อยู่ระหว่างการเรียบเรียง/เขียน (In preparation) ☐ ได้รับการจัดพิมพ์แล้ว (Published)
☑ 3.2.1.3 เอกสารประกอบการประชุม	☑ ก.ระดับชาติ (ระบุชื่อผู้วิจัย และ/หรือผู้ร่วมวิจัย ชื่อผลงานที่เสนอ ชื่อการประชุม วัน/เดือน/ปีที่จัด และสถานที่ ☑ Proceeding ☐ Book of Abstracts	
	☐ ข.ระดับนานาชาติ (ระบุชื่อผู้วิจัย และ/หรือผู้ร่วมวิจัย ชื่อผลงานที่เสนอ ชื่อการประชุม วัน/เดือน/ปีที่จัด และสถานที่ ☐ Proceeding ☐ Book of Abstracts	
3.3 การประชุมวิชาการ	☑ ก. ระดับชาติ (ระบุชื่อผู้วิจัย และ/หรือผู้ร่วมวิจัย ชื่อผลงานที่เสนอ ชื่อการประชุม วัน/เดือน/ปีที่จัด และสถานที่จัด) ☐ บรรยาย ☑ โปสเตอร์	
	☐ ข. ระดับนานาชาติ (ระบุชื่อผู้วิจัย และ/หรือผู้ร่วมวิจัย ชื่อผลงานที่เสนอ ชื่อการ	

ผลงาน	รายละเอียด
	ประชุม วัน/เดือน/ปีที่จัด และสถานที่จัด เมือง ประเทศ) 1. การประชุมในประเทศ ❑ บรรยาย ❑ โปสเตอร์ 2. การประชุมในต่างประเทศ ❑ บรรยาย ❑ โปสเตอร์
4. รางวัล/เกียรติบัตรที่ได้รับจากผลงานวิจัยนี้	
❑ ยังไม่เคยได้รับรางวัล/เกียรติบัตร	
☑ ได้รับรางวัล/เกียรติบัตร ดังนี้	
☑ ในประเทศ	เกียรติบัตรเพื่อรับรองว่าผลงานวิจัยเรื่อง การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล ได้ผ่านการพิจารณาจากคณะกรรมการผู้ทรงคุณวุฒิและได้นำเสนอผลงานประเภท Poster Presentation กลุ่มการวิจัยวิทยาศาสตร์และเทคโนโลยี (ด้านวิทยาศาสตร์ 1) วันพฤหัสบดีที่ 24 มกราคม 2562 ณ หอประชุมพญางำเมือง มหาวิทยาลัยพะเยา
❑ ต่างประเทศ	(ระบุชื่อรางวัล/เกียรติบัตรที่ได้รับ ผลงานที่ทำให้ได้รับรางวัล หน่วยงานที่มอบรางวัล ประเทศ และวัน/เดือน/ปีที่ได้รับ)

กิตติกรรมประกาศ

งานวิจัยฉบับนี้ได้รับทุนจากกองทุนพัฒนาการวิจัยและบริหารจัดการงานวิจัย มหาวิทยาลัยราชภัฏพิบูลสงคราม ทุนวิจัยเพื่อพัฒนานักวิจัย ประเภททุนนักวิจัยรุ่นใหม่ ประจำปีงบประมาณ พ.ศ. 2561 ผู้วิจัยจึงขอขอบพระคุณเป็นอย่างสูงต่อการสนับสนุนโครงการวิจัยเรื่องการพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล มา ณ ที่นี้

ผู้วิจัยหวังเป็นอย่างยิ่งว่าผลการวิจัยนี้จะเป็นประโยชน์แก่บุคลากรทางการศึกษาและผู้สนใจทั่วไป ตลอดจนจะเป็นประโยชน์ในการนำไปประยุกต์ใช้ต่อไป

ชฎารัตน์ ถาปัน
สลิลทิพย์ แดงกองโค
กุมภาพันธ์ 2562

หัวข้องานวิจัยเรื่อง (ภาษาไทย) การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่าง แบบชุดลำดับได้ดุล

คำสำคัญ การสุ่มตัวอย่าง ตัวแบบถดถอยเชิงเส้น การสุ่มตัวอย่างแบบชุดลำดับ

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling: BRSS) และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งได้ดำเนินการจำลองข้อมูลประชากรตามแผนการจำลองข้อมูล จากนั้นทำการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อสร้างตัวแบบถดถอย ซึ่งเกณฑ์ที่ใช้ในการพิจารณาตัวแบบถดถอยคือ ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2) และค่าความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) ผลการศึกษาโดยสรุปพบว่า ตัวแบบถดถอยที่สร้างจากตัวอย่างที่ได้มาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เมื่อมีขนาดตัวอย่างเพิ่มมากขึ้นสำหรับการสุ่มตัวอย่าง 1 รอบ ($m=1$) ตัวแบบถดถอยมีค่าสัมประสิทธิ์การตัดสินใจสูงที่สุด และมีค่าความคลาดเคลื่อนกำลังสองของตัวแบบถดถอยน้อยที่สุด

Research Title Development of Simple Linear Model by Using Balance Ranked Set Sampling.

Keywords sampling, linear regression model, ranked set sampling

ABSTRACT

The purposes of this study are to propose simple linear regression model based on balance ranked set sampling and to check the accuracy and reliability of simple linear regression model with balance ranked set sampling. The investigation was done mainly via simulation study. Finite population was generated by simulation plans. Then, ranked set sampling was applied so as to obtain sample data to create a model. The criteria use for consideration of simple linear regression model are coefficient of determination (R^2) and mean square error (MSE). It was found that the simple linear regression model was sampling by balance ranked set sampling when increasing sample size for cycle of sampling at 1 cycle ($m=1$) to obtain a most coefficient of determination and at least MSE.

สารบัญเรื่อง

หน้า

สารบัญเรื่อง (ต่อ)

สารบัญตาราง

หน้า

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

ในการจัดทำข้อสรุปสำหรับประชากรอันตะ หรือประชากรจำกัด (Finite Population) เมื่อย้อนกลับไปในช่วงแรกเริ่มนั้นระเบียบวิธีที่ใช้ในการจัดทำข้อสรุปของประชากร คือการสำมะโน (Census) อันได้แก่การเก็บรวบรวมข้อมูลทุกหน่วยประชากร แล้วนำข้อมูลนั้นมาคำนวณค่าลักษณะประชากร (Population Characteristics) เช่น ค่าเฉลี่ยประชากร (Population Average) ค่ายอดรวมประชากร (Population Total) ค่าสัดส่วนประชากร (Population Proportion) หรือค่าอัตราส่วนประชากร (Population Ratio) ด้วยข้อจำกัดของการสำมะโนไม่ว่าจะเป็นในส่วนของความทันสมัยของข้อมูล เวลาในการดำเนินการ บุคลากร และงบประมาณในการเก็บข้อมูล ต่อมาได้มีการพัฒนาระเบียบวิธีการเก็บรวบรวมข้อมูลที่เรียกว่า การสำรวจตัวอย่าง โดยทำการเก็บรวบรวมข้อมูลเพียงบางส่วนของประชากร ซึ่งในยุคแรกเริ่มของการสำรวจตัวอย่างนั้นจะทำการสำรวจโดยแผนการสุ่มตัวอย่างในลักษณะที่ไม่อิงความน่าจะเป็น (Non-probabilistic Sampling) [ชฎารัตน์ ถาปัน, 2555] เนื่องด้วยแผนการสุ่มตัวอย่างที่ไม่อิงความน่าจะเป็นมีข้อจำกัดในส่วนของความน่าเชื่อถือได้ ต่อมาจึงได้มีการพัฒนาแผนการสุ่มตัวอย่างในแนวทางที่อิงความน่าจะเป็น (Probabilistic Sampling) จึงได้มีการพัฒนาเป็นทฤษฎีการสำรวจตัวอย่าง (Theory of Sample Survey) แผนการสุ่มตัวอย่างที่อิงความน่าจะเป็นประเภทแรก คือ แผนการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) แต่แผนการสุ่มตัวอย่างแบบง่ายเกิดข้อจำกัดทั้งในส่วนของความผิดพลาดในการประมาณค่าลักษณะประชากรที่สูง รวมทั้งข้อจำกัดในการบริหารจัดการเก็บข้อมูล (Data Collection) หรือเรื่องเวลาและค่าใช้จ่าย ต่อมาจึงได้มีแผนการสุ่มตัวอย่างที่ถูกพัฒนาขึ้นมาเพื่อแก้ปัญหาที่เกิดขึ้น อาทิเช่น แผนการสุ่มตัวอย่างแบบมีชั้นภูมิ (Stratified Sampling) และแผนการสุ่มตัวอย่างแบบกลุ่ม (Cluster Sampling) เป็นต้น

แผนการสุ่มตัวอย่างดังกล่าวนั้นสามารถประมาณค่าลักษณะประชากรได้ หากแต่มีข้อจำกัดแตกต่างกันออกไป ซึ่งอาจส่งผลต่อความน่าเชื่อถือได้ ในปี 1952 McIntyre ได้เสนอวิธีการสุ่มตัวอย่างแบบชุดลำดับ (Ranked Set Sampling: RSS) เพื่อแก้ปัญหาการได้มาของแต่ละหน่วยตัวอย่างที่มีต้องใช้เวลานาน และมีค่าใช้จ่ายสูง ซึ่ง McIntyre ได้ใช้แผนการสุ่มตัวอย่างนี้ในการประมาณค่าเฉลี่ยของปริมาณหญ้าในพื้นที่หนึ่ง ซึ่งต้องใช้เวลาในการเก็บเกี่ยวทั้งหมด จากนั้นต้องใช้เวลาในการตากแห้งแล้วจึงทำการชั่งน้ำหนัก ปัญหาดังกล่าวนี้ McIntyre ได้นำมาแก้ปัญหาการเก็บข้อมูล [McIntyre G.A, 1952]

แผนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลมีวิธีการดำเนินการโดยสุ่มตัวอย่างมา k ชุดตัวอย่าง ชุดตัวอย่างละ k หน่วย แล้วทำการเรียงลำดับหน่วยตัวอย่างในแต่ละชุดตัวอย่าง จากนั้นจึงทำการเลือกตัวอย่างจากทั้ง k ชุดตัวอย่าง มาเพียงชุดตัวอย่างละ 1 หน่วย โดยชุดตัวอย่างที่ 1 เลือกหน่วยตัวอย่างลำดับที่ 1 จนกระทั่งชุดตัวอย่างที่ k เลือกหน่วยตัวอย่างที่ k ซึ่งจะทำการสุ่มเช่นนี้ไปจนครบ m รอบ แผนการสุ่มตัวอย่างนี้ทำให้ลดขนาดตัวอย่างในการเก็บรวบรวมข้อมูล และได้ค่าประมาณลักษณะประชากรที่น่าเชื่อถือได้เมื่อเปรียบเทียบกับแผนการสุ่มตัวอย่างแบบง่าย ดังการศึกษาของ Dell และ Clutter [Dell and Clutter, 1972]

ในทางปฏิบัติมีการนำข้อมูลที่ได้จากการสำรวจตัวอย่างไปหาค่าประมาณลักษณะประชากรแล้ว นอกจากนี้ยังมีการนำข้อมูลที่เก็บรวบรวมไปทำการวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูล รวมไปถึงการ

สร้างตัวแบบเพื่อพยากรณ์ ซึ่งตัวแบบที่นิยมใช้กันอย่างแพร่หลายคือ ตัวแบบถดถอยเชิงเส้น (Linear Regression Model) ซึ่งแบ่งออกเป็นตัวแบบถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Model) และตัวแบบถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Model)

ตัวแบบถดถอยเชิงเส้นอย่างง่ายเป็นตัวแบบที่วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระ 1 ตัวแปร กับตัวแปรตาม 1 ตัวแปร โดยตัวแปรทั้งสองมีความสัมพันธ์กันเชิงเส้นตรง ซึ่งมีตัวแบบถดถอยเชิงเส้นอย่างง่ายของประชากร คือ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ มี X_i เป็นตัวแปรอิสระ Y_i เป็นตัวแปรตาม β_0, β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า และ ε_i เป็นความคลาดเคลื่อน ในการวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล และเพื่อตรวจสอบความคลาดเคลื่อนของตัวประมาณค่าพารามิเตอร์ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

1.2 วัตถุประสงค์ของโครงการ

1. เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล
2. เพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล

1.3 ขอบเขตของโครงการวิจัย

1.3.1 แผนการจำลองข้อมูล

วัตถุประสงค์การศึกษาของโครงการวิจัยนี้เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล และตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล มีแผนการจำลองข้อมูลดังนี้

1. ประชากรที่ศึกษาสร้างขึ้นมาภายใต้การแจกแจงปกติ (Normal Distribution) โดยกำหนดให้ข้อมูลประชากรมีความสัมพันธ์กัน นั่นคือข้อมูลประชากรตัวแปรอิสระ (x) และตัวแปรตาม (y) มีความสัมพันธ์กันเชิงเส้นตรง
2. กำหนดค่าสหสัมพันธ์ ρ (Correlations) สำหรับตัวแปรอิสระและตัวแปรตาม แบ่งออกเป็น 3 ระดับความสัมพันธ์ ได้แก่
    1) ระดับความสัมพันธ์มาก คือมีค่า ρ อยู่ในช่วง 0.71-1.00
    2) ระดับความสัมพันธ์ปานกลาง คือมีค่า ρ อยู่ในช่วง 0.31-0.70
    3) ระดับความสัมพันธ์ต่ำ คือมีค่า ρ อยู่ในช่วง 0.00-0.30 [Ratner, 2009]
3. ขนาดประชากร (N) ที่ศึกษาเท่ากับ 1,000
4. ขนาดของตัวอย่าง (n) ที่ศึกษา คือ $n = 5, 10, 30, 40, 50$ [Voorhis and Morgan, 2007] โดยที่ $n = mk$ ซึ่ง k คือ ขนาดตัวอย่างในการสุ่มแต่ละชุดตัวอย่างและเป็นจำนวนชุดตัวอย่างของการสุ่มแต่ละรอบ และ m คือ จำนวนรอบในการสุ่มตัวอย่าง
5. ทำการทดลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ที่กำหนด (จำนวนการทดลอง 1,000 ครั้ง ในแต่ละสถานการณ์ เป็นจำนวนการทดลองที่มากพอในการจำลองข้อมูล ดังในการศึกษาการจำลองข้อมูลของ Alexander Levis และ John A. Sokolowsak and Catherine M. Banks) [Levis, 2017] [Sokolowsak and Banks, 2009]

จากแผนการจำลองข้อมูล จะทำการจำลองข้อมูลประชากรแล้วทำการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เพื่อทำการสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย เพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่าย

1.3.2 การตรวจสอบความถูกต้องและความน่าเชื่อถือได้

การตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล พิจารณาจากค่าสัมประสิทธิ์การตัดสินใจ และค่าความคลาดเคลื่อนกำลังสอง ดังนี้ [Douglas and Geoffrey, 2012]

1. ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination)

$$R^2 = \left(\frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y_i\right)\right)^2}} \right)^2$$

2. ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error)

$$MSE\left(\hat{Y}\right) = E\left(\hat{Y} - Y\right)^2$$

1.4 ประโยชน์ที่ได้รับจากการศึกษา

1. ได้ตัวแบบถดถอยเชิงเส้นถดถอยอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล
2. เป็นแนวทางเพื่อศึกษาการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

บทที่ 2
ตัวแบบถดถอยเชิงเส้นอย่างง่ายจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับประชากรอันตะ

2.1 ความนำ

การจัดทำข้อสรุปสำหรับประชากรอันตะ หรือประชากรจำกัด (Finite Population) ซึ่งสามารถทำการนับหน่วยประชากรได้อย่างครบถ้วนนั้น โดยพื้นฐานแล้วระเบียบวิธีที่ใช้คือ การสำมะโน (Census) แต่การสำมะโนประชากรนี้ยังมีข้อจำกัดอยู่หลายประการ สำหรับในกรณีที่ประชากรมีขนาดใหญ่ และหน่วยประชากรกระจายออกไปตามสภาพทางภูมิศาสตร์ อันได้แก่ เวลา บุคคลากร และงบประมาณเป็นจำนวนมากในการดำเนินการสำมะโนแต่ละครั้ง

ด้วยข้อจำกัดข้างต้น ต่อมาจึงได้มีการพัฒนาระเบียบวิธีการเก็บรวบรวมข้อมูลที่เรียกว่า การสำรวจตัวอย่าง โดยทำการเก็บรวบรวมข้อมูลเพียงบางส่วนของประชากร ต่อมาได้มีการพัฒนาวิธีการสำรวจตัวอย่างที่อิงความน่าจะเป็น และได้พัฒนาเป็นทฤษฎีการสำรวจตัวอย่าง (Theory of Sample Survey) เพื่อทำการอนุมานในอันที่จะอธิบายสภาพความเป็นไปเกี่ยวกับประชากรภายใต้กรอบการศึกษาในเชิงพรรณนา (Descriptive Inference) โดยการประมาณค่าลักษณะประชากร (Population Characteristics) อันได้แก่ ยอดรวมประชากร ค่าเฉลี่ยประชากร สัดส่วนประชากร และอัตราส่วนประชากร เป็นเบื้องต้นดัง William G. Cochran ได้กล่าวไว้ใน Sampling Techniques [Cochran, 1977]

ในระยะเวลาต่อมา เมื่อต้องการอนุมานหรือจัดทำข้อสรุปเกี่ยวกับประชากรอันตะในเชิงวิเคราะห์ (Analytical Inference) จึงได้มีการนำแนวคิดสำหรับการจัดทำข้อสรุปสำหรับประชากรอนันต์มาประยุกต์ใช้ เช่น การสร้างตัวแบบทางสถิติ (Statistical Modeling) เป็นต้น [ยงยุทธ ไชยพงศ์, 2552]

2.2 การอนุมานทางสถิติสำหรับประชากรอันตะ

2.2.1 ประชากรอันตะ (Finite Population)

ความหมายของคำว่าประชากรสำหรับในความเข้าใจสำหรับบุคคลทั่วไป หรือในสาขาวิชาอื่น ๆ นั้น อาจหมายความเพียงบุคคล หรือกลุ่มคนเท่านั้น แต่สำหรับในทางสถิติศาสตร์นั้น ประชากรที่ศึกษาไม่ได้หมายถึงเฉพาะกลุ่มคนเท่านั้น แต่ประชากรในทางสถิติศาสตร์ คือ กลุ่มคน สัตว์ สิ่งของ วัตถุ กลไก ผลผลิตทางการเกษตร หรืออะไรก็แล้วแต่ที่อยู่ภายใต้กรอบความสนใจ

ซึ่งถ้าหากประชากรที่ศึกษานั้นเกิดขึ้นภายใต้สถานการณ์ หรือกลไกเชิงสุ่มด้วยฟังก์ชันความน่าจะเป็นเดียวกันแล้วนั้น ภายใต้ทฤษฎีการอนุมานทางสถิติสำหรับประชากรอนันต์ เราก็จะสามารถอธิบายประชากรที่ศึกษาโดยการอธิบายกลไกเชิงสุ่มด้วยฟังก์ชันความน่าจะเป็นที่ทำให้เกิดประชากร หากแต่ตามสภาพความเป็นจริงแล้วนั้นยังมีประชากรบางประเภทที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา จนไม่สามารถอธิบายประชากรด้วยกลไกเชิงสุ่มด้วยฟังก์ชันความน่าจะเป็นใด ๆ ได้อย่างเหมาะสม และมีประสิทธิภาพ [ยงยุทธ ไชยพงศ์, 2552]

แนวทางแก้ปัญหาหนึ่งในการอธิบายประชากรที่มีลักษณะดังกล่าวนี้ ดำเนินการโดยกำหนดกรอบเวลา (Sampling Frame) ให้กับประชากรในขณะที่ทำการศึกษา ซึ่งเมื่อทำการกำหนดกรอบเวลาแล้วประชากรที่สนใจศึกษาเหล่านี้ก็จะสามารถนับจำนวนหน่วยประชากรได้อย่างครบถ้วน จึงทำให้สามารถอธิบาย

ลักษณะประชากรที่ศึกษาได้อย่างชัดเจนยิ่งขึ้นสำหรับในช่วงเวลาที่ได้กำหนดขึ้นนั้น ในทางสถิติศาสตร์เรียกประชากรในลักษณะเช่นนี้ว่าประชากรอันตะ ดังนั้นภายใต้แนวคิดเช่นนี้ ประชากรอันตะในทางสถิติศาสตร์จึงเป็นประชากรที่สามารถนับจำนวนหน่วยประชากรได้อย่างครบถ้วน คือมีขนาด N หน่วย โดยคุณลักษณะของประชากรที่ต้องการศึกษานั้นเป็นค่าคงที่ ที่ติดอยู่กับแต่ละหน่วยประชากร และไม่สามารถเปลี่ยนแปลงได้ [Cochran, 1977]

2.2.2 ทฤษฎีการสำรวจตัวอย่าง (Theory of Sample Survey)

ทฤษฎีการสำรวจตัวอย่างเป็นทฤษฎีที่ว่าด้วยการศึกษาแนวคิดในการหาข้อสรุปสำหรับประชากรอันตะโดยการสำรวจตัวอย่างที่อิงความน่าจะเป็นซึ่งทำการการเก็บข้อมูลเพียงบางส่วนจากประชากร หรือการสุ่มหน่วยประชากรขึ้นมาจำนวนหนึ่ง จากจำนวนประชากร N หน่วย แล้วนำข้อมูลที่ได้ไปทำการประมาณค่าลักษณะประชากรเพื่ออธิบายประชากรทั้งหมด ซึ่งเป็นที่แน่นอนว่า ค่าลักษณะประชากรที่ได้จากประชากรเพียงบางส่วนนั้นย่อมมีความแตกต่างจากค่าที่ได้จากสำมะโน หรือค่าจริงของประชากร จึงทำให้เกิดความผิดพลาดขึ้น

ความผิดพลาดที่เกิดขึ้นนั้นอาจมาจากสาเหตุ 2 ประการ คือ เกิดจากการเก็บรวบรวมข้อมูลเพียงบางส่วนของประชากร จึงทำให้ค่าที่ได้ไม่สามารถอธิบายประชากรได้อย่างถูกต้อง ซึ่งความผิดพลาดนี้เรียกว่า ความผิดพลาดจากการสุ่มตัวอย่าง (Sampling Error) และความผิดพลาดอีกสาเหตุหนึ่งอาจเกิดขึ้นตอนการเก็บรวบรวมข้อมูล การบันทึกข้อมูล และการประมวลผล เรียกความผิดพลาดนี้ว่า ความผิดพลาดที่ไม่ได้เกิดจากการสุ่มตัวอย่าง (Non-sampling Error)

การสำรวจตัวอย่างในแนวทางที่อิงความน่าจะเป็นนั้นเป็นที่ยอมรับ เนื่องจากสามารถควบคุมความผิดพลาดที่เกิดจากการสุ่มตัวอย่าง และประเมินระดับความผิดพลาดของค่าประมาณลักษณะประชากรจากข้อมูลตัวอย่าง และใช้ขนาดตัวอย่างน้อยกว่าการสุ่มตัวอย่างแบบอัตตะวิสัยจึงได้พัฒนาเป็นทฤษฎีการสำรวจตัวอย่างในที่สุด [Cochran, 1997]

2.2.3 การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling)

วิธีการสำรวจตัวอย่างที่อิงความน่าจะเป็นที่ถูกพัฒนาขึ้นมาแรกสุดคือ การสุ่มตัวอย่างแบบง่ายในการหาข้อสรุปสำหรับประชากรอันตะนั้น การสุ่มตัวอย่างแบบง่ายเป็นรากฐานสำคัญที่สุดสำหรับทฤษฎีการสำรวจตัวอย่าง ซึ่งเมื่อมีการนิยามประชากรที่สนใจแล้ว กระบวนการสุ่มตัวอย่างแบบง่ายเริ่มจากการสุ่มหน่วยประชากรขึ้นมาทีละหน่วย จนครบจำนวน n หน่วย จากประชากรทั้งหมด N หน่วย

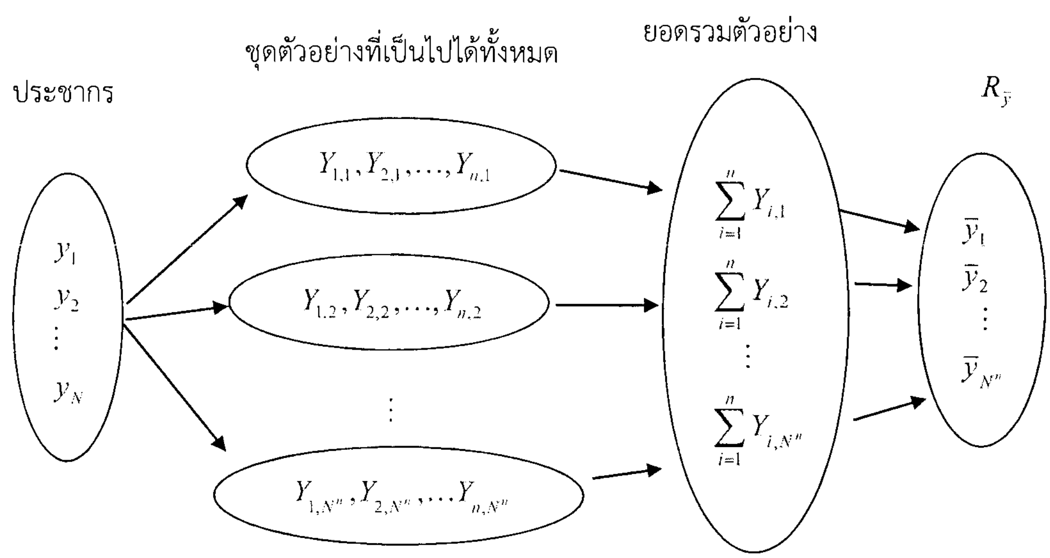
การสุ่มตัวอย่างแบบง่ายแบ่งได้ 2 ลักษณะ คือการสุ่มตัวอย่างแบบง่ายใส่คืน (Simple Random Sampling with Replacement: SRSWR) และการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน (Simple Random Sampling without Replacement: SRSWOR) ซึ่งการสุ่มตัวอย่าง 2 ลักษณะนี้ ทำให้ข้อมูลตัวอย่างที่ได้มีความน่าจะเป็นสำหรับการได้หน่วยตัวอย่างแต่ละหน่วย และฟังก์ชันความน่าจะเป็นของการเกิดขึ้นของชุดตัวอย่างต่างกัน ดังนั้นจึงส่งผลให้ตัวประมาณค่าลักษณะประชากรที่ได้มีความแปรปรวนต่างกันด้วย

การสุ่มตัวอย่างแบบง่ายใส่คืนมีกระบวนการสุ่มตัวอย่างเริ่มจาก สุ่มหน่วยประชากรขึ้นมาเป็นตัวอย่างครั้งละ 1 หน่วย เป็นจำนวนทั้งหมด n ครั้ง จากประชากรทั้งหมด N หน่วย ซึ่งมีเงื่อนไขของการสุ่ม คือ ในแต่ละครั้งของการสุ่มตัวอย่างนั้น หน่วยประชากรใดที่ถูกสุ่มขึ้นมาเป็นตัวอย่างแล้วจะถูกนำกลับคืนไปยัง

ประชากรอีก นั่นคือ หน่วยประชากรใดที่ถูกสุ่มขึ้นมาแล้วจะมีโอกาสถูกสุ่มเป็นตัวอย่างในครั้งต่อ ๆ ไปได้ หรือหน่วยประชากรหน่วยนั้นมีโอกาสถูกสุ่มมากกว่า 1 ครั้งด้วยความน่าจะเป็นที่เท่ากัน

การทดลองสุ่มสำหรับการสุ่มตัวอย่างแบบง่ายใส่คืน คือการสุ่มหน่วยประชากรขึ้นมา 1 หน่วย จากประชากรทั้งหมด N หน่วย ซึ่งประกอบด้วย n การทดลองสุ่มย่อย โดยที่แต่ละครั้งของการทดลองสุ่มนั้นมีผลลัพธ์ที่เป็นไปได้ N วิธี และผลลัพธ์ที่ได้จากการสุ่มแต่ละครั้งไม่มีผลกระทบต่อการสุ่มตัวอย่างครั้งต่อไป จึงทำให้การทดลองสุ่มย่อยทั้งหมด n การทดลองสุ่มมีคุณสมบัติที่เป็นอิสระต่อกัน ดังนั้นการนับจำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมดจึงได้ใช้หลักการคูณ (Multiplication Principle) นั่นคือมีจำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมด N^n ชุดตัวอย่าง [ยงยุทธ ไชยพงศ์, 2552]

จากกระบวนการสุ่มตัวอย่างในแต่ละครั้งของการสุ่มตัวอย่าง ผลลัพธ์ที่ได้ไม่มีผลกระทบต่อการสุ่มตัวอย่างในครั้งต่อไป ดังนั้นความน่าจะเป็นของการได้หน่วยตัวอย่างแต่ละหน่วยจึงเกิดขึ้นด้วยความน่าจะเป็นที่เท่ากัน ภายใต้แนวคิดแบบคลาสสิคจะได้ความน่าจะเป็นของการได้หน่วยตัวอย่างในแต่ละครั้งของการสุ่มตัวอย่างคือ $\frac{1}{N}$ และแต่ละชุดตัวอย่าง $(Y_1, Y_2, \ldots, Y_n)$ เกิดขึ้นด้วยความน่าจะเป็น $P(Y_1, Y_2, \ldots, Y_n) = \frac{1}{N^n}$ ซึ่งรูปแบบความน่าจะเป็นที่กำกับการเกิดขึ้นของชุดตัวอย่างนี้จะนำไปสู่รูปแบบความน่าจะเป็นของตัวประมาณ คือ ยอดรวมตัวอย่าง และค่าเฉลี่ยตัวอย่าง สำหรับการสุ่มตัวอย่างแบบใส่คืน [ยงยุทธ ไชยพงศ์, 2552]



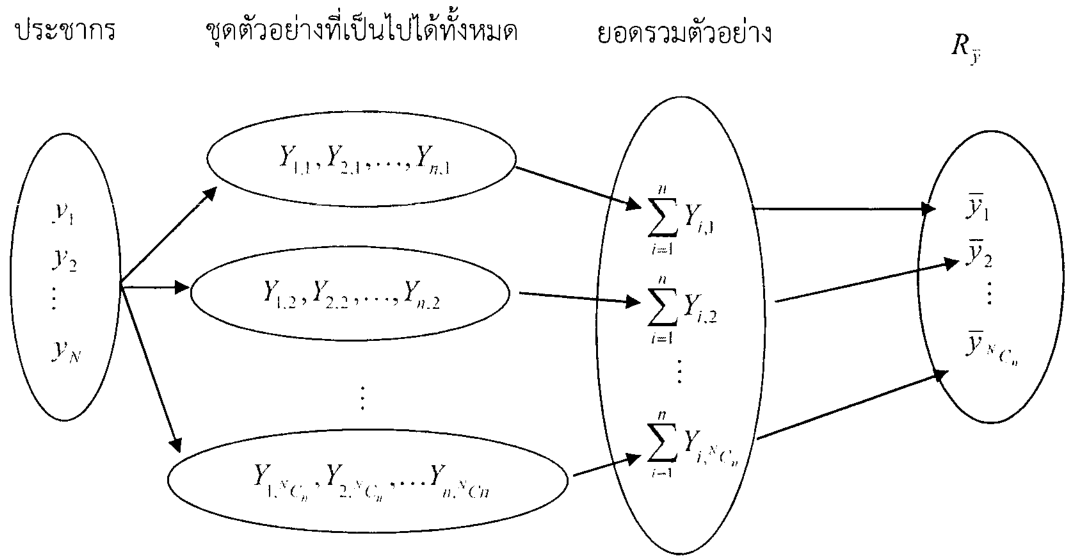
ภาพที่ 2.1 ชุดตัวอย่างที่เป็นไปได้ทั้งหมดและการเกิดขึ้นของตัวแปรสุ่มยอดรวมตัวอย่าง และค่าเฉลี่ยตัวอย่างภายใต้แผนการสุ่มตัวอย่างแบบง่ายที่ใส่คืน

สำหรับการสุ่มตัวอย่างแบบง่ายไม่ใส่คืนมีกระบวนการสุ่มตัวอย่าง คือ สุ่มหน่วยประชากรขึ้นมา n หน่วย จากประชากรทั้งหมด N หน่วย ซึ่งจะทำการสุ่มตัวอย่างมาครั้งละ 1 หน่วย โดยมีเงื่อนไขว่าหน่วยประชากรใดที่ถูกสุ่มขึ้นมาเป็นตัวอย่างแล้วจะไม่นำกลับไปใส่คืนยังประชากรอีก ดังนั้นการสุ่มตัวอย่างใน

ครั้งแรกจะมีผลต่อการสุ่มตัวอย่างในครั้งถัดไป และหน่วยประชากรใดที่ยังไม่ถูกสุ่มเป็นตัวอย่างมีโอกาสถูกสุ่มเป็นตัวอย่างที่เท่ากัน

การทดลองสุ่มที่เกิดขึ้นสำหรับการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน คือการสุ่มหน่วยประชากรขึ้นมาเป็นตัวอย่างครั้งละ 1 หน่วย จนครบ n ครั้ง โดยที่แต่ละครั้งของการสุ่มมีผลต่อการสุ่มตัวอย่างในครั้งถัดไป นั่นคือ การสุ่มตัวอย่างในครั้งแรกมีผลลัพธ์ที่เป็นไปได้ทั้งหมด N วิธี ส่วนครั้งที่สอง มีผลลัพธ์ที่เป็นไปได้ทั้งหมด $N-1$ วิธี และจนกระทั่งครั้งที่ n มีผลลัพธ์ที่เป็นไปได้ทั้งหมด $N-(n-1)$ วิธี นั่นคือการทดลองสุ่มสำหรับการสุ่มตัวอย่างแบบง่ายไม่ใส่คืนทั้ง n ครั้งไม่เป็นอิสระต่อกัน ดังนั้นการนับจำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมดจะใช้หลักการจัดกลุ่มที่ลำดับไม่มีความสำคัญ (Combination) นั่นคือ จำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมด คือ ${}^{N}C_{n}$ ชุด [Cochran, 1997]

จากกระบวนการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน ผลลัพธ์ของการได้หน่วยตัวอย่างแรกมีผลกระทบต่อการสุ่มตัวอย่างครั้งต่อไป ดังนั้นความน่าจะเป็นสำหรับการได้หน่วยตัวอย่างที่ 1 คือ $\frac{1}{N}$ และความน่าจะเป็นของการได้หน่วยตัวอย่างหน่วยที่ 2 คือ $\frac{1}{N-1}$ จนกระทั่งหน่วยตัวอย่างที่ n มีความน่าจะเป็น คือ $\frac{1}{N-(n-1)}$ เนื่องจากหน่วยประชากรที่ถูกสุ่มมาแล้วจะไม่ถูกสุ่มมาเป็นตัวอย่างซ้ำอีก และจากจำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมดคือ ${}^{N}C_{n}$ ชุด แต่ละชุดตัวอย่างจึงเกิดขึ้นด้วยความน่าจะเป็น $P(Y_1, Y_2, \ldots, Y_n) = 1/{}^{N}C_{n}$ ภายใต้กรอบแนวคิดของความน่าจะเป็นแบบคลาสสิก



ภาพที่ 2.2 ชุดตัวอย่างที่เป็นไปได้ทั้งหมด และการเกิดขึ้นของตัวแปรสุ่มยอดรวมตัวอย่างและค่าเฉลี่ยตัวอย่างภายใต้แผนการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน

2.3 การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling: BRSS)

ในการสำรวจตัวอย่างเพื่อเก็บรวบรวมข้อมูลในกรณีที่ประชากรที่ศึกษามีขนาดใหญ่จึงทำให้ต้องใช้เวลา และค่าใช้จ่ายสำหรับการศึกษาเป็นจำนวนมาก ในปี 1972 McIntyre จึงได้พัฒนาวิธีการสุ่มตัวอย่างขึ้นมาเพื่อหาวิธีการสุ่มตัวอย่างที่สามารถลดค่าใช้จ่าย (Cost-effective Sampling Method) และประหยัดเวลาในการดำเนินการ ซึ่งวิธีการสุ่มตัวอย่างดังกล่าวคือ วิธีการสุ่มตัวอย่างแบบชุดลำดับได้ดุล [McIntyre, 1952]

วิธีการสุ่มตัวอย่างแบบชุดลำดับได้ดุลได้พัฒนามาจากแผนการสุ่มตัวอย่างแบบง่ายซึ่งแผนการสุ่มตัวอย่างแบบง่ายนั้นมีข้อจำกัดด้านกรอบตัวอย่าง หรือบัญชีรายชื่อหน่วยประชากร นั่นคือเมื่อต้องทำการสุ่มตัวอย่างแบบง่ายจะต้องมีกรอบตัวอย่าง หรือบัญชีรายชื่อหน่วยประชากรจึงจะสามารถดำเนินการเก็บรวบรวมข้อมูลได้ ส่วนแผนการสุ่มตัวอย่างแบบมีชั้นภูมิที่ได้พัฒนามาจากแผนการสุ่มตัวอย่างแบบง่ายนั้น เป็นแผนการสุ่มตัวอย่างที่แบ่งประชากรขนาด N ออกเป็นชั้นภูมิ แล้วจึงทำการสุ่มตัวอย่างในแต่ละชั้นภูมิ ซึ่งแผนการสุ่มตัวอย่างแบบมีชั้นภูมินี้ก็ต้องมีกรอบตัวอย่าง หรือบัญชีรายชื่อเช่นเดียวกับแผนการสุ่มตัวอย่างแบบง่าย

ต่อมาเมื่อมีความต้องการศึกษาประชากรที่มีขนาดใหญ่มาก ๆ จึงได้พัฒนาเป็นแผนการสุ่มตัวอย่างแบบกลุ่มขึ้น โดยทำการแบ่งประชากรออกเป็นกลุ่มย่อยตามภูมิศาสตร์ แล้วเก็บข้อมูลตัวอย่างจากกลุ่มที่แบ่งซึ่งไม่จำเป็นต้องมีรายชื่อทุกหน่วยประชากร แต่การสุ่มตัวอย่างแบบกลุ่มนี้ก็ยังมีข้อจำกัดด้านความเชื่อถือได้ของค่าประมาณ ซึ่งอยู่ในระดับต่ำกว่าแผนการสุ่มตัวอย่างแบบง่าย

และเพื่อตอบสนองความต้องการในขณะนั้น จึงได้หาวิธีที่สามารถลดค่าใช้จ่าย และเวลาในการดำเนินการเก็บรวบรวมข้อมูล เพื่อใช้ในการประมาณปริมาณหญ้าของทุ่งหญ้าในประเทศออสเตรเลีย เมื่อต้นปี 1952 McIntyre จึงได้นำเสนอวิธีการสุ่มตัวอย่างขึ้นมาวิธีหนึ่ง ซึ่งต่อมาภายหลัง Hall และDell ได้เรียกวิธีการสุ่มตัวอย่างดังกล่าวว่า การสุ่มตัวอย่างแบบชุดลำดับ (Ranked Set Sampling) [Halls and Dell, 1966]

แผนการสุ่มตัวอย่างของ McIntyre ดำเนินการสุ่มตัวอย่างดังนี้ เริ่มจากการแบ่งทุ่งหญ้าออก เป็นล็อต (Lots) แล้วทำการสุ่มตัวอย่างทุ่งหญ้าที่ถูกแบ่งเป็นล็อตมา k ล็อต จากนั้นทำการเรียงลำดับแต่ละล็อตตามปริมาณหญ้าด้วยการคาดคะเน หรือการประมาณด้วยสายตา (Eye Inspection) ซึ่งจะทำการสุ่มตัวอย่างในลักษณะดังกล่าวมา k ครั้ง

จากนั้นทำการเลือกหน่วยตัวอย่าง ซึ่งตัวอย่างจากการสุ่มครั้งที่ 1 เลือกหญ้าในล็อตลำดับที่ 1 เพื่อทำการตัด และนำไปชั่งน้ำหนัก ส่วนตัวอย่างจากการสุ่มครั้งที่ 2 เลือกหญ้าในล็อตลำดับที่ 2 แล้วดำเนินการเช่นเดียวกันกับล็อตลำดับที่ 1 และดำเนินการเช่นนี้ไปเรื่อย ๆ จนกระทั่งครบ k ครั้ง ซึ่งจะได้ตัวอย่างหญ้า k ล็อต การสุ่มตัวอย่างดังกล่าว เรียกว่าการสุ่มตัวอย่าง 1 รอบ (Cycle) และหากต้องการจำนวนตัวอย่างที่เพิ่มขึ้นสามารถทำการสุ่มตัวอย่างตามกระบวนการดังกล่าวมา m รอบ จากนั้นจึงทำการประมาณปริมาณหญ้า [McIntyre, 1952]

การสุ่มตัวอย่างแบบชุดลำดับนั้นมีขั้นตอนการสุ่มตัวอย่างที่ซับซ้อนมากกว่าการสุ่มตัวอย่างแบบง่ายนั่นคือการสุ่มตัวอย่างแบบง่ายจะทำการสุ่มตัวอย่างขึ้นมาเพียง n หน่วยเท่านั้นจากประชากรขนาด N หน่วย แต่สำหรับการสุ่มตัวอย่างแบบชุดลำดับจะทำการสุ่มตัวอย่างมา n_r หน่วย และ $r = 1,2,\ldots,k$ เรียกว่าชุดตัวอย่าง ซึ่งจะทำการสุ่มตัวอย่างขึ้นมา k ชุดตัวอย่าง เรียกว่าการสุ่มตัวอย่าง 1 รอบ และกำหนดให้ $Y_{[r]i'}$ คือ หน่วยตัวอย่างลำดับที่ r ชุดตัวอย่างที่ i' สำหรับการสุ่มตัวอย่าง 1 รอบ ได้แสดงดังนี้

$$
\begin{array}{l}
Y_{[1]1}\ Y_{[1]2}\ \cdots\ Y_{[1]n_1} \\
Y_{[2]1}\ Y_{[2]2}\ \cdots\ Y_{[2]n_2} \\
\cdots\ \ \cdots\ \ \cdots\ \ \cdots \\
Y_{[k]1}\ Y_{[k]2}\ \cdots\ Y_{[k]n_k}
\end{array}
$$

ถ้าขนาดตัวอย่างในการสุ่มแต่ละชุดตัวอย่างเท่ากัน คือ $n_1 = n_2 = \ldots = n_k$ แล้วจะเรียกการสุ่มตัวอย่างแบบชุดลำดับนี้ว่า การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling) และถ้าขนาดตัวอย่างที่สุ่มมาในแต่ละชุดตัวอย่างมีขนาดไม่เท่ากันแล้ว จะเรียกการสุ่มตัวอย่างนี้ว่า การสุ่มตัวอย่างแบบชุดลำดับไม่ได้ดุล (Unbalance Ranked Set Sampling)

2.3.1 ขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

สำหรับขั้นตอนของการสุ่มตัวอย่างแบบชุดลำดับได้ดุลนั้น มีขั้นตอนการสุ่มตัวอย่างเริ่มต้นด้วยการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน (SRSWOR) ขนาด k หน่วย จากประชากรขนาด N หน่วย ซึ่งจะเรียกว่าชุดตัวอย่างที่ 1 แล้วนำหน่วยตัวอย่าง (Sampling Units) ทั้ง k หน่วยมาเรียงลำดับจากน้อยไปหามากด้วยการคาดคะเน หรือการประมาณด้วยสายตา (Judgment Ranking) ด้วยตัวแปรที่สนใจ จากนั้นหน่วยตัวอย่างที่อยู่ในลำดับที่ 1 จะถูกเลือกเป็นตัวอย่าง และทำการวัดค่าของตัวแปรที่สนใจ ส่วนหน่วยตัวอย่างที่เหลือถูกตัดทิ้งหรือไม่สนใจ

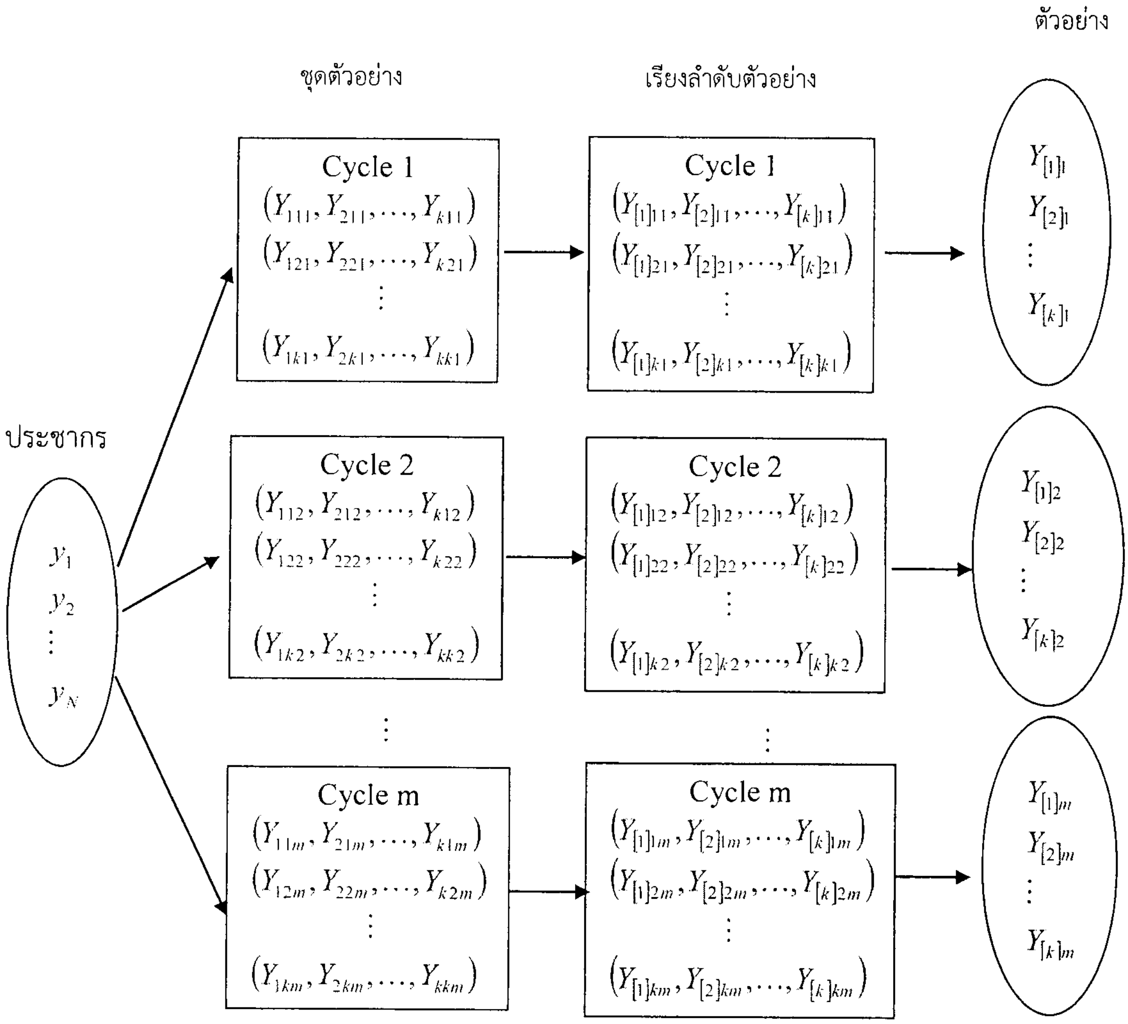
เมื่อสิ้นสุดการสุ่มตัวอย่างชุดที่ 1 แล้ว ต่อจากนั้นสุ่มตัวอย่างแบบง่ายไม่ใส่คืนขนาด k หน่วย จากประชากรขนาด N หน่วย ซึ่งนั่นหมายความว่า หน่วยประชากรที่ถูกสุ่มเป็นตัวอย่างในชุดตัวอย่างที่ 1 ถูกนำกลับมาใส่ในประชากรอีกครั้ง หน่วยประชากร k หน่วยที่ถูกสุ่มในครั้งที่ 2 เรียกว่าชุดตัวอย่างที่ 2 จากนั้นนำหน่วยตัวอย่างมาเรียงลำดับจากน้อยไปหามาก หน่วยตัวอย่างที่อยู่ในลำดับที่ 2 ในชุดตัวอย่างที่ 2 นี้จะถูกเลือกเป็นตัวอย่าง แล้วทำการวัดค่าของตัวแปรที่สนใจ และหน่วยที่เหลือถูกตัดทิ้งหรือไม่สนใจ

กระบวนการสุ่มตัวอย่างนี้ดำเนินการต่อไปเรื่อย ๆ จนกระทั่งสุ่มตัวอย่างไปจนครบ k ชุดตัวอย่าง และเรียกกระบวนการสุ่มตัวอย่างจนครบ k ชุดตัวอย่างนี้ว่า “รอบ (Cycle)” ซึ่งจะทำการสุ่มตัวอย่างทั้งหมด m รอบ และการสุ่มหน่วยตัวอย่างในแต่ละชุดตัวอย่างที่ทำการสุ่มขึ้นมาทั้ง k ชุดตัวอย่างในแต่ละรอบนั้น การได้หน่วยตัวอย่าง k หน่วย ในแต่ละชุดตัวอย่างไม่มีผลกระทบต่อการได้หน่วยตัวอย่างในชุดตัวอย่างอื่น ๆ

นั่นคือ ชุดตัวอย่างที่ 1 จะให้ตัวอย่างหน่วยที่ 1 ชุดตัวอย่างที่ 2 จะให้ตัวอย่างหน่วยที่ 2 จนถึงชุดตัวอย่างที่ k จะให้หน่วยตัวอย่างที่ k เมื่อเสร็จสิ้นการสุ่มตัวอย่างในรอบที่ 1 จะได้ตัวอย่าง k หน่วย และจากนั้นจึงทำการสุ่มตัวอย่างรอบที่ 2 ซึ่งก็จะได้ตัวอย่าง k หน่วย และเมื่อทำการสุ่มตัวอย่างไปเรื่อย ๆ จนถึงรอบที่ m ซึ่งได้ตัวอย่างมา k หน่วยเช่นเดิม ดังนั้นข้อมูลตัวอย่างที่เกิดขึ้นทั้งหมดจึงมี mk หน่วย

เนื่องจากในการสุ่มตัวอย่างในแต่ละรอบและแต่ละชุดตัวอย่างจะสุ่มจากประชากรที่มีขนาด N หน่วย ทุกครั้ง ดังนั้นหน่วยประชากรใดที่ถูกสุ่มมาเป็นตัวอย่างแล้วยังสามารถถูกสุ่มขึ้นมาเป็นหน่วยตัวอย่างในชุดตัวอย่างต่อไปได้สำหรับการสุ่มตัวอย่างครั้งต่อไป ดังนั้นชุดตัวอย่างทั้ง k ชุดตัวอย่าง (ในแต่ละรอบ) เป็นอิสระต่อกัน และเกิดขึ้นด้วยความน่าจะเป็นเท่ากัน

เมื่อชุดตัวอย่างทั้ง k ชุดตัวอย่าง เป็นอิสระกันแล้วการสุ่มตัวอย่างแต่ละรอบย่อมเป็นอิสระต่อกัน และได้ตัวอย่างสำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุลที่มีขนาดเท่ากับ mk หน่วย [Kowalczyk, 2004] ซึ่งขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลได้แสดงดังภาพที่ 2.3



หมายเหตุ $Y_{[r]i}$ คือ หน่วยตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งอยู่ในลำดับที่ r รอบที่ i ; $r = 1,2,3,\ldots,k$ และ $i = 1,2,3,\ldots,m$

ภาพที่ 2.3 แสดงขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

จากภาพที่ 2.3 เป็นการสุ่มตัวอย่างแบบชุดลำดับได้ดุลมา m รอบ รอบละ k ชุดตัวอย่าง ได้ตัวอย่างคือ $Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[k]1}, \ldots, Y_{[1]m}, Y_{[2]m}, \ldots, Y_{[k]m}$ มีขนาด mk หน่วย ซึ่งหน่วยตัวอย่างแรกคือ $Y_{[1]1}$ ได้จากชุดตัวอย่างที่ 1 ของการสุ่มตัวอย่างรอบที่ 1 และหน่วยตัวอย่างสุดท้ายคือ $Y_{[k]m}$ ซึ่งได้จากชุดตัวอย่างที่ k ของการสุ่มตัวอย่างรอบที่ m [ชฎารัตน์ ถาปัน, 2555]

2.3.2 คุณสมบัติข้อมูลตัวอย่างของการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

การสุ่มตัวอย่างแบบชุดลำดับได้ดุลนั้น ข้อมูลตัวอย่างได้มาจากการสุ่มหน่วยประชากรเพียงบางส่วนขึ้นมาเป็นตัวอย่าง และหากต้องการทราบว่าข้อมูลตัวอย่างที่ได้นั้นมีคุณสมบัติเป็นอย่างไรจึงต้องทำการวิเคราะห์ถึงขั้นตอนการสุ่มตัวอย่าง การทดลองสุ่มที่เกิดขึ้น และการเรียงลำดับหน่วยตัวอย่าง

ขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลประกอบด้วย 2 ขั้นตอน คือการสุ่มตัวอย่าง และการเรียงลำดับหน่วยตัวอย่าง ซึ่งการสุ่มตัวอย่างเพื่อให้ได้ขนาดตัวอย่าง mk หน่วย นั้นจะต้องทำการสุ่ม

ตัวอย่างมา m รอบ ซึ่งในแต่ละรอบของการสุ่มนั้นจะสุ่มมา k ชุดตัวอย่าง นั่นคือต้องทำการสุ่มตัวอย่างมาทั้งหมด mk ชุดตัวอย่าง

เมื่อพิจารณาถึงการสุ่มตัวอย่างรอบที่ 1 คือการสุ่มหน่วยประชากรขนาด k หน่วย จำนวน k ชุดตัวอย่าง ซึ่งการสุ่มตัวอย่างในแต่ละชุดตัวอย่างจะสุ่มหน่วยประชากรขึ้นมาจำนวน k หน่วยนั้นเป็นการสุ่มตัวอย่างแบบง่ายไม่ใส่คืน การดำเนินการดังกล่าวจึงมีคุณสมบัติเป็นการทดลองสุ่ม เนื่องจากในการสุ่มตัวอย่างขึ้นมา k หน่วยนั้น เป็นการสุ่มหน่วยประชากรขึ้นมาทีละหน่วยจากประชากรขนาด N หน่วย โดยหน่วยประชากรที่ถูกสุ่มขึ้นมาแล้วจะไม่ถูกนำไปใส่คืนประชากร ซึ่งการดำเนินการดังกล่าวไม่สามารถทราบผลลัพธ์ได้ว่าหน่วยประชากรหน่วยใดจะถูกสุ่มขึ้นมาเป็นตัวอย่าง โดยมีผลลัพธ์ที่เป็นไปได้ทั้งหมดของการทดลองสุ่มถูกเก็บไว้ในปริภูมิตัวอย่าง

ซึ่งปริภูมิตัวอย่างสำหรับการสุ่มตัวอย่างหน่วยที่ 1 สำหรับชุดตัวอย่างที่ 1 คือ $S_1 = \{y_1, y_2, \ldots, y_N\}$ และให้ E_1 เป็นเหตุการณ์ของการได้มาซึ่งหน่วยตัวอย่างที่ 1 ดังนั้น $E_1 = \{y_{i_1}\}$ โดยที่ $i_1 = 1,2,\ldots,N$ จากกระบวนการสุ่มตัวอย่างดังกล่าว หน่วยประชากรแต่ละหน่วยถูกสุ่มขึ้นมาเป็นตัวอย่างด้วยความน่าจะเป็นที่เท่ากัน ดังนั้นแต่ละผลลัพธ์ในปริภูมิตัวอย่าง $S_1 = \{y_1, y_2, \ldots, y_N\}$ จึงเกิดขึ้นด้วยความน่าจะเป็นที่เท่ากัน (Equally Likely Outcomes) และความน่าจะเป็นที่เกิดขึ้นของหน่วยตัวอย่างที่ 1 คือ $\Pr(E_1) = \dfrac{1}{N}$

ส่วนการสุ่มตัวอย่างหน่วยที่ 2 สำหรับชุดตัวอย่างที่ 1 เนื่องจากมีหน่วยประชากร 1 หน่วยถูกสุ่มออกไปเป็นตัวอย่างแล้ว และไม่ถูกนำกลับคืนไปยังประชากร ดังนั้นปริภูมิตัวอย่างสำหรับการสุ่มตัวอย่างหน่วยที่ 2 คือ $S_2 = \{y_1, y_2, \ldots, y_{N-1}\}$ และให้ E_2 เป็นเหตุการณ์ของการได้มาซึ่งหน่วยตัวอย่างที่ 2 ดังนั้น $E_2 = \{y_{i_2}\}$ โดยที่ $i_2 = 1,2,\ldots,N-1$ ดังนั้นความน่าจะเป็นที่เกิดขึ้นของหน่วยตัวอย่างที่ 2 คือ $\Pr(E_2) = \dfrac{1}{N-1}$ และเมื่อทำการสุ่มตัวอย่างมาจนกระทั่งถึงหน่วยตัวอย่างที่ k ความน่าจะเป็นที่เกิดขึ้นของหน่วยตัวอย่างที่ k คือ $\Pr(E_k) = \dfrac{1}{N-(k-1)}$

และในการสุ่มตัวอย่างแบบชุดลำดับได้ดุลนั้นเมื่อสุ่มตัวอย่างมาได้ครบ k หน่วยแล้วนั้น ตัวอย่างทั้ง k หน่วย จะถูกนำมาเรียงลำดับจากน้อยไปหามาก ซึ่งการเรียงลำดับหน่วยตัวอย่างนี้เป็นการทดลองสุ่มที่ดำเนินการต่อเนื่องมาจากการสุ่มตัวอย่าง และการทดลองสุ่มที่เกิดขึ้นคือ หน่วยตัวอย่างทั้ง k หน่วยจะถูกเรียงลำดับ ซึ่งการเรียงลำดับหน่วยตัวอย่างนี้ไม่สามารถทราบมาก่อนว่าหน่วยตัวอย่างทั้ง k หน่วย หน่วยตัวอย่างใดจะอยู่ในลำดับที่เท่าไหร่ ดังนั้นการเรียงลำดับจึงเป็นการทดลองสุ่ม

เมื่อหน่วยตัวอย่างทั้ง k หน่วยถูกเรียงลำดับแล้วจะทำการเลือกหน่วยตัวอย่างซึ่งหน่วยตัวอย่างที่อยู่ในลำดับที่ 1 จะถูกเลือกเป็นตัวอย่างสำหรับชุดตัวอย่างที่ 1 นั่นคือจะมีผลลัพธ์ที่สามารถเป็นไปได้ทั้งหมดอยู่ k หน่วย สำหรับชุดตัวอย่างที่ 1 และมีปริภูมิตัวอย่าง คือ $\{y_1, y_2, \ldots, y_k\}$ และความน่าจะเป็นของหน่วยตัวอย่างที่อยู่ในลำดับที่ 1 ก็คือ สถิติอันดับ (Order Statistic) เนื่องจากมีการจัดเรียงลำดับข้อมูลตัวอย่าง

และสำหรับการสุ่มตัวอย่างของชุดตัวอย่างที่ 2 มีการดำเนินการสุ่มตัวอย่าง และการเรียงลำดับดำเนินการเช่นเดียวกันกับชุดตัวอย่างที่ 1 เนื่องจากว่าในการสุ่มตัวอย่างแต่ละชุดตัวอย่างเป็นอิสระต่อกันดังได้กล่าวไว้แล้วในขั้นตอนการสุ่มตัวอย่าง

ดังนั้นการสุ่มหน่วยประชากรแต่ละหน่วยจึงมีปริภูมิตัวอย่างเช่นเดียวกับชุดตัวอย่างที่ 1 ส่วนการเลือกหน่วยตัวอย่างในชุดตัวอย่างที่ 2 นั้น หน่วยตัวอย่างที่อยู่ในลำดับที่ 2 จะถูกเลือกเป็นตัวอย่าง จนกระทั่งในชุดตัวอย่างที่ k หน่วยตัวอย่างที่ k จะถูกเลือกเป็นตัวอย่าง ซึ่งความน่าจะเป็นของหน่วยตัวอย่างในแต่ละลำดับก็คือ สถิติอันดับที่ r นั่นเอง และหน่วยประชากรที่ถูกสุ่มขึ้นมาเป็นตัวอย่างที่อยู่ในลำดับที่ r จึงมีคุณสมบัติเป็นตัวแปรสุ่มนั่นเอง และการได้หน่วยตัวอย่างที่ 1 ($Y_{[1]1}$) ไม่มีผลกระทบต่อหน่วยตัวอย่างอื่น ๆ เนื่องจากว่าแต่ละชุดตัวอย่างถูกสุ่มมาจากประชากรขนาด N หน่วย เหมือนกัน [ชฎารัตน์ ถาปัน, 2555]

ในกรณีของการเลือกหน่วยตัวอย่าง ซึ่งหน่วยตัวอย่างที่เลือกนั้นมาจากการเรียงลำดับหน่วยตัวอย่างในแต่ละชุดตัวอย่าง ซึ่งเมื่อต้องการหาความน่าจะเป็นของหน่วยตัวอย่างแต่ละหน่วยจึงต้องใช้สถิติอันดับมาอธิบาย ซึ่งมีนิยามดังนี้

นิยามที่ 1 กำหนดให้ $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n หน่วย สถิติอันดับ คือ การนำตัวแปรสุ่มนี้มาเรียงลำดับจากน้อยไปหามาก ซึ่งเขียนแทนด้วย $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$

ดังนั้น สถิติอันดับ คือตัวแปรสุ่มที่มีที่มีคุณสมบัติ $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ โดยที่ $X_{(1)}$ คือค่าของตัวแปรสุ่มที่น้อยที่สุดของทั้งชุดตัวแปรสุ่ม $X_1, X_2, \ldots, X_n$ และ $X_{(2)}$ คือค่าของตัวแปรสุ่มที่น้อยที่สุดเป็นอันดับสอง และ $X_{(n)}$ คือค่าที่มากที่สุดของทั้งชุดตัวแปรสุ่ม $X_1, X_2, \ldots, X_n$ [Casella and Berger, 2002]

สำหรับการหาความน่าจะเป็นของตัวแปรสุ่มลำดับที่ r ($Y_{[r]i}$) หรือความน่าจะเป็นของตัวอย่างในลำดับที่ r ภายใต้ประชากรอันตะ ในปี 1995 Patil, Sinha และ Taillie ได้ศึกษาสถิติอันดับจากประชากรอันตะ (Order Statistics from Finite Population) สำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งให้ $\Omega = \{y_1, y_2, \ldots, y_N\}$ คือประชากรอันตะที่มีขนาด N หน่วย และมีค่าเฉลี่ยคือ μ ความแปรปรวน คือ σ^2 และได้สมมติประชากรที่ศึกษามีลักษณะ $y_1 \le y_2 \le \ldots \le y_N$ เพื่อใช้ในการหาความน่าจะเป็นสำหรับหน่วยตัวอย่างลำดับที่ r ที่ถูกสุ่มมาจากประชากร ซึ่งเมื่อทำการสุ่มตัวอย่างขนาด k หน่วย แบบง่ายไม่ใส่คืนจากประชากร Ω ได้กำหนดเหตุการณ์ที่สนใจคือ

$$\{r \Rightarrow s\}$$

จากข้างต้น $\{r \Rightarrow s\}$ คือ หน่วยตัวอย่างลำดับที่ r ที่ได้มาจากหน่วยประชากรในลำดับที่ s ซึ่งความน่าจะเป็นที่จะได้หน่วยตัวอย่างลำดับที่ r คือ

$$A_r^s = \Pr\{r \Rightarrow s\}$$

และความน่าจะเป็นสำหรับหน่วยตัวอย่างในลำดับที่ r คือ

$$A_r^s = \frac{\binom{s-1}{r-1}\binom{N-s}{k-r}}{\binom{N}{k}}, \quad s = 1,2,\ldots,N$$

จากความน่าจะเป็น A_r^s เป็นการหาความน่าจะเป็นภายใต้แนวคิดของประชากรอันตะ ความน่าจะเป็นของหน่วยตัวอย่างในเหตุการณ์ $\{r \Rightarrow s\}$ หาจากการสุ่มตัวอย่างขนาด k หน่วย จากประชากรขนาด N หน่วย ซึ่งในตัวอย่าง k หน่วยที่สุ่มมาเมื่อเรียงลำดับแล้วจะสนใจหน่วยตัวอย่างในลำดับที่ r ซึ่ง $r-1$ ถูกเลือกจากประชากร $s-1$ และ $k-r$ หน่วยถูกเลือกมาจาก ประชากร $N-s$ [Patil, Sinha and Taillie, 1995]

จากขั้นตอนการสุ่มตัวอย่าง ผลที่ได้รับคือ ตัวอย่างที่มีขนาด mk หน่วย ดังนี้

$$Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[k]1}, \ldots\ , Y_{[1]m}, Y_{[2]m}, \ldots, Y_{[k]m}$$

ซึ่งตัวอย่างถูกสุ่มมาจากประชากรขนาด N หน่วย ซึ่งสามารถนับจำนวนหน่วยประชากรได้อย่างครบถ้วน และค่าคุณลักษณะของแต่ละหน่วยประชากรเป็นข้อมูลเชิงปริมาณซึ่งเป็นค่าคงที่ที่ติดอยู่กับหน่วยประชากรแต่ละหน่วย ซึ่งค่าเหล่านั้นคือตัวแปรที่มีค่าคงที่ (Fixed Variable) แต่เมื่อทำการสุ่มหน่วยประชากรขึ้นมาเป็นตัวอย่างโดยที่ไม่ทราบว่าหน่วยประชากรหน่วยใดจะถูกสุ่มขึ้นมาเป็นตัวอย่าง ดังนั้นค่าคุณลักษณะของประชากรที่ตกเป็นตัวอย่างจึงเป็นตัวแปรสุ่ม ซึ่งปริภูมิมิติของตัวแปรสุ่มมีจำนวนสมาชิกที่นับได้อย่างครบถ้วน ดังนั้นตัวแปรสุ่มลำดับที่ r ($Y_{[r]i}$) เป็นตัวแปรสุ่มแบบไม่ต่อเนื่อง

ค่าคาดหวัง สำหรับตัวแปรสุ่ม $Y_{[r]i}$ หาได้ดังนี้

$$\begin{aligned} E\left(Y_{[r]i}\right) &= \sum_{s=1}^{N} y_s \Pr\left(Y_{[r]i} = y_s\right) \\ &= \sum_{s=1}^{N} y_s \Pr\left(\{i \Rightarrow s\}\right) = \sum_{s=1}^{N} y_s A_i^s \end{aligned}$$

และในทำนองเดียวกันความแปรปรวนสำหรับตัวแปรสุ่ม $Y_{[r]i}$ หาได้ดังนี้

$$\begin{aligned} Var\left(Y_{[r]i}\right) &= E\left(Y_{[r]i}\right)^2 - \left(E\left(Y_{[r]i}\right)\right)^2 \\ &= \sum_{s=1}^{N} Y_s^2 A_i^s - \left(\sum_{s=1}^{N} y_s A_i^s\right)^2 \end{aligned}$$

[Patil, Sinha and Taillie, 1995]

2.3.3 ความน่าจะเป็นของการเกิดขึ้นของชุดตัวอย่างสำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

การสุ่มตัวอย่างแบบชุดลำดับได้ดุลเป็นการสุ่มตัวอย่างขึ้นมาชุดตัวอย่างละ k หน่วย ซึ่งทำการสุ่มทั้งหมด k ชุดตัวอย่าง ซึ่งหมายถึงการสุ่มตัวอย่าง 1 รอบ และจะทำการสุ่มตัวอย่างทั้งหมด m รอบ ดังนั้นจึงมีชุดตัวอย่างที่ทำการสุ่มตัวอย่างทั้งหมด คือ mk ชุดตัวอย่าง

การนับจำนวนชุดตัวอย่างภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับประชากรอันตะนั้น ถ้าหากพิจารณาการสุ่มตัวอย่างแบบชุดลำดับได้ดุลในแต่ละชุดตัวอย่าง ซึ่งเป็นการสุ่มแบบง่ายไม่ใส่คืน กระบวนการสุ่มตัวอย่างนี้มีคุณสมบัติเป็นการทดลองสุ่มที่ประกอบไปด้วย k การทดลองสุ่ม ดังนั้นการนับจำนวนชุดตัวอย่างที่เป็นไปได้ทั้งหมดสำหรับการสุ่มตัวอย่างมาครั้งละ k หน่วย คือ ${}^{N}C_k$ ชุดตัวอย่าง ดังแผนภาพที่ 2.5

ดังนั้นภายใต้กรอบแนวคิดของความน่าจะเป็นแบบคลาสสิก (Classical Probability) ชุดตัวอย่างที่ 1 เกิดขึ้นด้วยความน่าจะเป็น $\frac{1}{{}^{N}C_k}$ และ ชุดตัวอย่างที่ 2 เกิดขึ้นด้วยความน่าจะเป็น $\frac{1}{{}^{N}C_k}$ ดังเช่นชุดตัวอย่างที่ 1 เนื่องจากการสุ่มชุดตัวอย่างในแต่ละชุดตัวอย่างเป็นอิสระต่อกัน หรือการสุ่มตัวอย่างในชุดตัวอย่างที่ 1 ไม่มีผลต่อการสุ่มตัวอย่างในชุดตัวอย่างอื่นๆ ดังนั้นความน่าจะเป็นสำหรับของการเกิดขึ้นของชุดตัวอย่างสำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุล คือ $\frac{1}{{}^{N}C_k}$ [ชฎารัตน์ ถาปัน, 2555]

2.3.4 ตัวประมาณค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

สำหรับการอนุมานทางสถิติสำหรับประชากรอนันต์ วิธีการสร้างตัวประมาณนั้นมีอยู่มากมายหลายวิธี วิธีการแบบโมเมนต์เป็นวิธีการพื้นฐานในการสร้างตัวประมาณแบบจุด ซึ่งวิธีการนี้จะสร้างตัวประมาณจากโมเมนต์ตัวอย่างรอบจุดกำเนิด และโมเมนต์ของประชากรรอบจุดกำเนิด จากนั้นตัวประมาณแบบจุดจะได้จากสมการโมเมนต์ตัวอย่างรอบจุดกำเนิดเท่ากับโมเมนต์ประชากรรอบจุดกำเนิด แต่ยังมีข้อจำกัดในการสร้างตัวประมาณ คือต้องสามารถหาโมเมนต์ของประชากรให้ได้

วิธีการสร้างตัวประมาณแบบจุดที่ถูกยอมรับกันอย่างแพร่หลายคือ วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) แต่สำหรับในส่วนของประชากรอันตะนั้นยังมีข้อจำกัดในการนำวิธีภาวะน่าจะเป็นสูงสุดไปสร้างตัวประมาณ ยกตัวอย่างเช่น กรณีของการสุ่มตัวอย่างแบบง่ายในการประมาณค่าเฉลี่ยประชากร เมื่อทำการสร้างฟังก์ชันภาวะน่าจะเป็นสำหรับตัวอย่างสุ่มภายใต้แผนการสุ่มตัวอย่างแบบง่ายจะพบว่าฟังก์ชันภาวะน่าจะเป็นทั้งในกรณีที่ใส่คืน และไม่ใส่คืนไม่ขึ้นอยู่กับค่าลักษณะของประชากร และข้อมูลตัวอย่างเลย ดังนั้นการสร้างตัวประมาณด้วยวิธีการแบบภาวะน่าจะเป็นสูงสุดจึงไม่สามารถดำเนินการได้ เนื่องจากฟังก์ชันภาวะน่าจะเป็นไม่ได้เก็บสาระ (Information) เกี่ยวกับพารามิเตอร์ที่สนใจไว้ ซึ่งไม่สอดคล้องกับหลักการภาวะน่าจะเป็น (Likelihood Principle) จึงต้องใช้วิธีการสร้างตัวประมาณแบบโมเมนต์ในการสร้างตัวประมาณ หรือวิธีการที่ใช้ Empirical Likelihood Function

สำหรับการสร้างตัวประมาณค่าเฉลี่ยประชากรภายใต้แผนการสุ่มตัวอย่างแบบชุดลำดับได้ดุล วิธีที่ใช้ในการสร้างตัวประมาณค่าเฉลี่ยประชากร คือ วิธีโมเมนต์ ซึ่งได้ตัวประมาณค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลดังนี้

ในการสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งเมื่อทำการสุ่มตัวอย่างมา m รอบ แล้วจะได้จำนวนชุดตัวอย่างทั้งหมด mk ชุดตัวอย่าง ซึ่งมีขั้นตอนดำเนินการสุ่มตัวอย่างดังหัวข้อ 2.3.1 จะได้ตัวอย่างที่อยู่ในลำดับที่ 1 ทั้ง m รอบ คือ $Y_{[1]1}, Y_{[1]2}, Y_{[1]3}, \dots, Y_{[1]m}$ และหน่วยตัวอย่างในลำดับที่ 2 ทั้ง m รอบ คือ $Y_{[2]1}, Y_{[2]2}, Y_{[2]3}, \dots, Y_{[2]m}$ จนกระทั่งหน่วยตัวอย่างที่อยู่ในลำดับที่ k สำหรับการสุ่มตัวอย่าง m รอบ คือ $Y_{[k]1}, Y_{[k]2}, Y_{[k]3}, \dots, Y_{[k]m}$ ซึ่งได้ตัวประมาณค่าเฉลี่ยประชากร คือ $\hat{\bar{Y}}_{RSS} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k} Y_{[r]i} = \bar{y}_{RSS}$ [Patil, Sinha and Taillie, 1995]

2.3.5 คุณสมบัติของตัวประมาณค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

ในทางทฤษฎีการสำรวจตัวอย่างนั้น การตรวจสอบคุณสมบัติของตัวประมาณค่าลักษณะประชากรทำได้โดยใช้หลักการเดียวกันกับในทางทฤษฎีการอนุมานทางสถิติสำหรับประชากรอนันต์ทุกประการ แต่ต่างกันเพียงรูปแบบความน่าจะเป็นของตัวประมาณซึ่งเป็นรูปแบบที่เกิดจากการสุ่มหน่วยประชากรขึ้นมาเป็นตัวอย่าง แทนที่จะเป็นรูปแบบการแจกแจงร่วมของตัวอย่างสุ่มที่ถูกสร้างขึ้นมาจากรูปแบบความน่าจะเป็นที่กำกับกลไกก่อกำเนิดของหน่วยประชากร และการตรวจสอบคุณสมบัติของตัวประมาณค่าลักษณะประชากรซึ่งโดยพื้นฐานเป็นการหาค่าคาดหวัง และความแปรปรวนของตัวประมาณที่ถูกสร้างขึ้นมาสำหรับแผนการสุ่มตัวอย่างแต่ละประเภท

คุณสมบัติที่สำคัญของตัวประมาณคือ ความไม่เอนเอียง (Unbiased) ซึ่งจะแสดงให้ทราบว่าโดยเฉลี่ยแล้วตัวประมาณนั้น ๆ ไม่มีแนวโน้มให้ค่าประมาณสูงหรือต่ำกว่าค่าคุณลักษณะประชากร ตัวประมาณ $\bar{y}_{RSS}$ เป็นตัวประมาณค่าเฉลี่ยประชากรที่ไม่เอนเอียง สำหรับ $\bar{Y}$ ซึ่งการพิจารณาคุณสมบัติความไม่

เอนเอียงของตัวประมาณ $\bar{y}_{RSS}$ สามารถดำเนินการโดยใช้ค่าคาดหวังสำหรับแต่ละหน่วยตัวอย่าง $Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[k]1}, \ldots, Y_{[1]m}, Y_{[2]m}, \ldots, Y_{[k]m}$ ซึ่งได้ทำการพิสูจน์โดย Patil, Sinha และ Taillie ดังนี้

$$E(\bar{y}_{RSS}) = \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k} E\left[Y_{[r]i}\right]$$

$$= \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\sum_{s=1}^{N} A_i^s y_s = \frac{1}{N}\sum_{s=1}^{N} y_s = \bar{Y}$$

[Patil, Sinha and Taillie, 1995]

2.3.6 เปรียบเทียบคุณสมบัติตัวประมาณค่าเฉลี่ยภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลกับการสุ่มตัวอย่างแบบง่าย

ในปี 1980 Stokes ได้ศึกษาการประมาณค่าความแปรปรวนของตัวอย่างที่ได้มาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งความแปรปรวนของตัวอย่างนี้ได้มาจากวิธีโมเมนต์ (Moment Estimates) โดยศึกษาจากโมเมนต์ของสถิติลำดับ (Moments of Judgment Order Statistics)

ความแปรปรวนตัวอย่าง $Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[k]1}, \ldots, Y_{[1]m}, Y_{[2]m}, \ldots, Y_{[k]m}$ สำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุล คือ $s_{RSS}^2 = \sum_{i=1}^{m}\sum_{r=1}^{k}\left(Y_{[r]i} - \bar{y}_{RSS}\right)^2 \Big/ (mk-1)$ [Stokes, 1980]

การเปรียบเทียบคุณสมบัติของตัวประมาณ นอกจากการพิจารณาถึงคุณสมบัติความไม่เอนเอียงแล้ว ยังมีการพิจารณาถึงความแปรปรวนของตัวประมาณด้วย ซึ่งค่าความแปรปรวนของตัวประมาณเป็นค่าที่แสดงให้ทราบว่ากำลังสองของผลต่างระหว่างค่าประมาณกับค่าที่แท้จริงโดยเฉลี่ยแล้วเป็นเท่าใด นั่นคือหากตัวประมาณตัวใด มีค่าความแปรปรวนน้อยกว่าตัวประมาณตัวอื่น ๆ แล้ว ตัวประมาณนั้นย่อยมีคุณสมบัติเป็นตัวประมาณที่มีค่าความแปรปรวนต่ำสุด

การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าเฉลี่ยระหว่างการสุ่มตัวอย่างแบบชุดลำดับได้ดุลกับการสุ่มตัวอย่างแบบง่าย สามารถทำได้โดยการหาอัตราส่วนระหว่างความแปรปรวนของตัวประมาณที่ต้องการเปรียบเทียบดังนี้

$$RE(\bar{y}_{RSS}, \bar{y}_{SRS}) = \frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{RSS})}$$

จากอัตราส่วนระหว่างความแปรปรวนของตัวประมาณข้างต้น ถ้าหาก $RE(\bar{y}_{RSS}, \bar{y}_{SRS}) \geq 1$ แล้ว แสดงว่าตัวประมาณค่าเฉลี่ยภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลมีความแปรปรวนน้อยกว่าตัวประมาณค่าเฉลี่ยภายใต้การสุ่มตัวอย่างแบบง่าย [Chen, Bai and Sinha, 2003]

การเปรียบเทียบประสิทธิภาพตัวประมาณค่าเฉลี่ยระหว่างการสุ่มตัวอย่างแบบชุดลำดับได้ดุลกับการสุ่มตัวอย่างแบบง่ายจะทำการเปรียบเทียบกรณีการสุ่มตัวอย่างแบบง่ายใส่คืนสำหรับขนาดตัวอย่างเท่ากัน เนื่องจาก ในการสุ่มตัวอย่างแต่ละชุดตัวอย่างหรือแต่ละรอบนั้นสุ่มมาจากประชากร N หน่วย ซึ่งในที่นี้ขนาดตัวอย่าง $n = mk$ ดังนี้

$$RE(\bar{y}_{RSS}, \bar{y}_{SRS}) = \frac{\left(\dfrac{N-1}{N}\right)\dfrac{s^2}{mk}}{\dfrac{\sum_{i=1}^{m}(\bar{y}_i - \bar{y}_{RSS})^2}{m(m-1)}} = \left(\frac{N-1}{N}\right)\left(\frac{\sum_{j=1}^{mk}\left(y_j - \bar{y}_{SRS}\right)^2 \Big/ mk-1}{\sum_{i=1}^{m}(\bar{y}_i - \bar{y}_{RSS})^2 \Big/ m-1}\right) \geq 1$$

จากการเปรียบเทียบประสิทธิภาพของตัวประมาณข้างต้นนี้ พบว่า $RE(\bar{y}_{RSS},\bar{y}_{SRS})\geq 1$ เนื่องจาก $\sum_{i=1}^{m}(\bar{y}_i-\bar{y}_{RSS})^2 \big/ m-1$ มีค่าน้อยกว่า $\sum_{j=1}^{mk}(y_j-\bar{y}_{SRS})^2 \big/ mk-1$ นั่นคือตัวประมาณค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลมีความแปรปรวนน้อยกว่า เมื่อเปรียบเทียบกับตัวประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างแบบง่ายใส่คืน

และเมื่อทำการเปรียบเทียบประสิทธิภาพตัวประมาณโดยใช้ความแปรปรวนภายใต้ข้อสมมติของ [Kowalczyk, 2004] กับตัวประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างแบบง่ายใส่คืนสำหรับในกรณีที่ขนาดตัวอย่างเท่ากัน ได้ดังนี้

$$RE(\bar{y}_{RSS},\bar{y}_{SRS})=\frac{\left(\frac{N-1}{N}\right)\frac{s^2}{mk}}{\frac{1}{mk}\left[\left(1-\frac{1}{N}\right)s^2-\frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\left(E\left(Y_{[r]i}\right)-\bar{Y}\right)^2\right]}\geq 1$$

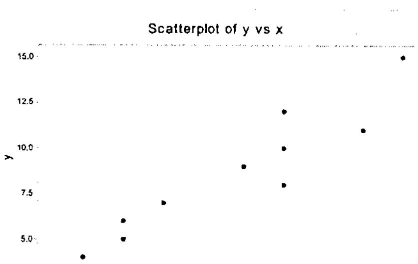
จากการเปรียบเทียบประสิทธิภาพของตัวประมาณข้างต้นนี้ พบว่า $RE(\bar{y}_{RSS},\bar{y}_{SRS})\geq 1$ เนื่องจาก $\left[\frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\left(E\left(Y_{[r]i}\right)-\bar{Y}\right)^2\right]\geq 0$ นั่นคือ ตัวประมาณโดยใช้ความแปรปรวนภายใต้ข้อสมมติของ [Kowalczyk, 2004] ภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลมีค่าความแปรปรวนน้อยกว่าเมื่อเปรียบเทียบกับตัวประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างแบบง่ายใส่คืน

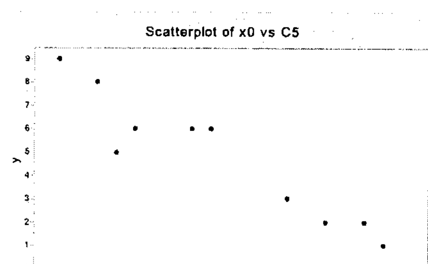
2.4 การวิเคราะห์ถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis)

การศึกษาข้อมูลซึ่งได้มาจากการเก็บรวบรวมหรือการสำรวจตัวอย่าง แล้วนำมาวิเคราะห์หาความสัมพันธ์เพื่อจัดทำข้อสรุป รวมทั้งการสร้างตัวแบบเพื่อการพยากรณ์นั้น เป็นวิธีที่นิยมใช้กันอย่างแพร่หลาย ไม่ว่าจะเป็นในเชิงธุรกิจ เศรษฐศาสตร์ เกษตรกรรม สังคมศาสตร์ รวมทั้งด้านการแพทย์ ถูกนำมาใช้เพื่อการพยากรณ์ถึงเหตุการณ์หรือข้อมูลในอนาคต

การวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเป็นการศึกษาความสัมพันธ์ระว่างตัวแปร 2 ตัว ซึ่งประกอบไปด้วยตัวแปรอิสระ (Independence Variable) 1 ตัว และตัวแปรตาม (Dependence Variable) 1 ตัว ซึ่งความสัมพันธ์ระหว่าง 2 ตัวแปรนี้เป็นไปในลักษณะความสัมพันธ์เชิงเส้นตรง

ความสัมพันธ์เชิงเส้นตรงนั้นจะพิจารณาจากแนวโน้มของค่าสังเกตซึ่งเมื่อนำมาเขียนเป็นแผนภาพการกระจาย (Scatter plot) ค่าสังเกตนั้นจะมีลักษณะหรือมีแนวโน้มเป็นเส้นตรง อาจจะมีความสัมพันธ์ตามกันหรือผกผันกันดังเช่นภาพที่2.4





ภาพที่2.4 แสดงแผนภาพการกระจายระหว่างตัวแปรอิสระ (x) และตัวแปรตาม (y)

2.4.1 ตัวแบบถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Model)

ในช่วงปลายศตวรรษที่ 18 Sri Francis Galton ได้ศึกษาความสัมพันธ์ระหว่างส่วนสูงของบุตรและส่วนสูงของพ่อแม่ ซึ่งพบว่าส่วนสูงของบุตรที่มีพ่อแม่สูงจะมีค่าเข้าใกล้ค่าเฉลี่ยประชาร ส่วนบุตรที่มีพ่อแม่เตี้ยมีแนวโน้มที่จะสูงขึ้นเพื่อเข้ามาใกล้ค่าเฉลี่ยของประชารด้วย Sri Francis Galton ได้อธิบายความสัมพันธ์ดังกล่าวด้วยคณิตศาสตร์ และต่อมาได้ใช้คำว่า "การถดถอย (Regression)" อธิบายความสัมพันธ์เชิงสถิติระหว่างตัวแปรต้นและตัวแปรตาม [พิษณุ เจียวคุณ, 2548]

ตัวแบบเชิงเส้นถดถอยอย่างง่ายดังที่ได้กล่าวมาแล้วข้างต้นว่า ตัวแบบนี้ประกอบไปด้วยตัวแปรอิสระ และตัวแปรตาม อย่างละ 1 ตัว โดยมีตัวแบบการถดถอยสำหรับประชากร (Population Regression Model) คือ

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

โดยที่ y_i เป็นค่าสังเกตของตัวแปรตาม

β_0 เป็นพารามิเตอร์แทนระยะตัดแกน Y

β_1 เป็นพารามิเตอร์แทนความชันของเส้นถดถอย

x_i เป็นค่าสังเกตของตัวแปรอิสระ (ค่าคงที่ที่ทราบค่า)

ε_i เป็นค่าความคลาดเคลื่อนเชิงสุ่ม

ในการนำตัวแบบถดถอยเชิงเส้นอย่างง่ายไปใช้ผู้ศึกษาควรที่จะทราบข้อสมมติที่เกี่ยวข้องกับตัวแบบถดถอยเชิงเส้นอย่างง่ายของประชากร ดังนี้

1. Y_i เป็นตัวแปรสุ่มที่มีการแจกแจงความน่าจะเป็น $f(Y_i \mid X_i)$ ที่มีค่าเฉลี่ย $\mu_{Y|X}$ และมีความแปรปรวนเท่ากับ $\sigma^2_{Y|X}$ ซึ่งมีค่าเท่ากันสำหรับทุก ๆ ค่าของ X_i

2. ค่าเฉลี่ยของ Y_i จะอยู่บนเส้นตรง ซึ่งเป็นเส้นถดถอยของประชากร นั่นคือ

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i \\ &= \mu_{Y|X} \end{aligned}$$

3. ค่าความคลาดเคลื่อน ε_i มีการแจกแจงปกติ มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ σ^2 นั่นคือ $\varepsilon_i \sim N(0, \sigma^2)$

4. ความคลาดเคลื่อน ε_i แต่ละค่าไม่มีความสัมพันธ์กัน นั่นคือ $Cov(\varepsilon_i, \varepsilon_j) = 0$ เมื่อ $i \neq j$

5. ค่าของ Y_i แต่ละค่าเป็นอิสระต่อกัน [พิษณุ เจียวคุณ, 2548]

สมบัติของเส้นถดถอย (Properties of Fitted regression line)

1. ผลรวมของค่าความคลาดเคลื่อนมีค่าเท่ากับศูนย์เสมอ ($\sum_{i=1}^{n} \varepsilon_i = 0$)

2. ผลรวมกำลังสองของค่าความคลาดเคลื่อน ($\sum_{i=1}^{n} \varepsilon_i^2$) จะมีค่าน้อยที่สุด

3. ผลรวมของค่าความคลาดเคลื่อนที่ถ่วงด้วยตัวแปรอิสระ X_i มีค่าเท่ากับศูนย์ ($\sum_{i=1}^{n} X_i \varepsilon_i = 0$)

4. ผลรวมของค่าความคลาดเคลื่อนที่ถ่วงด้วยค่าประมาณของ Y_i มีค่าเท่ากับศูนย์ ($\sum_{i=1}^{n} Y_i \varepsilon_i = 0$)

5. เส้นถดถอยจะผ่านจุด $\left(\overline{X}, \overline{Y}\right)$ เสมอ

[พิษณุ เจียวคุณ, 2548]

2.4.2 การประมาณค่าพารามิเตอร์ β_0 และ β_1 สำหรับตัวแบบถดถอยเชิงเส้นอย่างง่าย

ในทางปฏิบัติหรือสถานการณ์จริงเราไม่สามารถทำการเก็บรวบรวมข้อมูลประชากรได้อย่างครบถ้วน ด้วยข้อจำกัดไม่ว่าจะเป็นในเรื่องของการเก็บรวบรวมข้อมูล บุคคลากรในการดำเนินการ รวมไปถึงในเรื่องของงบประมาณในการเก็บรวมรวมข้อมูล ด้วยเหตุนี้จึงต้องมีการสำรวจตัวอย่าง เพื่อนำข้อมูลตัวอย่างหรือข้อมูลเพียงบางส่วนของประชากรมาเพื่อใช้ ทำให้ไม่สามารถทราบค่าพารามิเตอร์ β_0 และ β_1 ได้

ด้วยเหตุผลดังข้างต้นจึงต้องมีการประมาณค่าพารามิเตอร์ วิธีการประมาณค่าพารามิเตอร์ที่นิยมใช้สำหรับประมาณค่าพารามิเตอร์ β_0 และ β_1 คือวิธีกำลังสองน้อยสุด (Least Square Method) ซึ่งจะให้ตัวประมาณที่มีคุณสมบัติไม่เอนเอียง และความแปรปรวนต่ำสุด

หลักการหาตัวประมาณพารามิเตอร์ β_0 และ β_1 ของวิธีกำลังสองน้อยสุด มีหลักการคือทำให้ค่าผลรวมกำลังสองของส่วนเหลือหรือความคลาดเคลื่อน (Residual) มีค่าที่น้อยที่สุด กล่าวคือ ส่วนเหลือ (e_i) คือค่าความแตกต่างระหว่างค่าจริงของตัวแปรตาม y_i กับค่าพยากรณ์ที่ได้จากสมการถดถอย ($\hat{y}_i$) ที่ระดับเดียวกันของค่าของตัวแปรอิสระ X_i หรือสามารถเขียนได้ดังนี้

$$e_i = y_i - \hat{y}_i \qquad ; \ i = 1, 2, 3, \ldots, n$$

ซึ่งในที่นี้ตัวแบบถดถอยของประชากร คือ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ และตัวแบบถดถอยของตัวอย่าง (Sample Regression Model) คือ $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ กำหนดให้ Q เป็นค่าผลรวมกำลังสองของความคลาดเคลื่อนที่น้อยที่สุด ได้ดังนี้

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \ ; \quad i = 1, 2, 3, \ldots, n \qquad (2.1)$$

ตัวประมาณพารามิเตอร์ของ β_0 และ β_1 คือ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ตามลำดับ เพื่อให้ได้ค่า Q ที่น้อยที่สุดนั้นต้องหาอนุพันธ์ย่อยเทียบกับค่า β_0 และ β_1 ได้ดังนี้

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

และ

$$\left.\frac{\partial Q}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

ซึ่งจะได้สองสมการดังต่อไปนี้

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (2.2)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (2.3)$$

สมการที่ (2.2) และ (2.3) เราจะเรียกว่าสมการปกติกำลังสองน้อยสุด (Least-squares Normal Equation) ซึ่งเมื่อแก้ระบบสมการที่ (2.2) และ (2.3) จะได้ตัวประมาณดังนี้

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{2.4}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{2.5}$$

เมื่อ $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ และ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ [Douglas and Geoffrey, 2012]

นอกเหนือจากการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแล้ว ยังสามารถใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method)

หลักการทั่วไปของการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดจะพิจารณาจากการสร้างฟังก์ชันความหนาแน่นของค่าสังเกต Y_i ซึ่งฟังก์ชันความหนาแน่นนี้ได้มาจากเก็บรวบรวมข้อมูลหรือค่าสังเกตแล้วนำไปสร้าง Likelihood Function ซึ่งไม่ทราบค่าพารามิเตอร์ จากนั้นจะทำการหาตัวประมาณที่ทำให้ Likelihood Function หรือ $L(\theta)$ มีค่าสูงสุด

ซึ่งจากข้อสมมติของตัวแบบถดถอยเชิงเส้นอย่างง่ายของประชากรที่กล่าวว่า Y_i เป็นตัวแปรสุ่มที่มีการแจกแจงความน่าจะเป็น $f(Y_i \mid X_i)$ โดยมีค่าเฉลี่ย $E(Y_i) = \beta_0 + \beta_1 X_i$ และมีความแปรปรวน σ^2 จะได้ว่า Y_i มีการแจกแจงแบบปกติ ซึ่งหาตัวประมาณพารามิเตอร์ได้ดังนี้

จากตัวแบบถดถอยเชิงเส้นของประชากรคือ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ซึ่ง $\varepsilon_i \sim N(0, \sigma^2)$ จะได้ว่า ε_i มีรูปแบบการแจกแจงคือ $f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\varepsilon_i)^2}$ เมื่อ Transformation จะได้ตัวแปรสุ่ม Y_i มีการแจกแจงแบบปกติ ซึ่งมีรูปแบบการแจกแจงคือ $f(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_i))^2}$

เมื่อทำการสร้าง Likelihood Function

$$L(Y_i; \beta_0, \beta_i) = f(Y_1) f(Y_2) \ldots f(Y_n)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) \tag{2.6}$$

จากนั้นจะทำการหาอนุพันธ์ย่อยเทียบกับพารามิเตอร์ β_0 และ β_1 เพื่อหาตัวประมาณที่ทำให้ Likelihood Function มีค่าสูงสุด จะได้ว่า

$$\ln L(Y_i; \beta_0, \beta_i) = n \ln\left(2\pi\sigma^2\right)^{-\frac{1}{2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.7}$$

$$\frac{\partial}{\partial \beta_0}\left(\ln L(Y_i; \beta_0, \beta_i)\right) = -\frac{1}{2\sigma^2} 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \beta_1 \bar{x} \tag{2.8}$$

และ $$\frac{\partial}{\partial \beta_1}\left(\ln L\left(Y_i;\beta_0,\beta_i\right)\right) = -\frac{1}{2\sigma^2}\cdot 2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)\left(-x_i\right)=0$$

$$\beta_1 = \frac{n\left(\sum_{i=1}^{n}x_i y_i\right)-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2} \tag{2.9}$$

ให้ตัวประมาณของ β_0 และ β_1 คือ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ตามลำดับ จะได้ว่า

$$\hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} \tag{2.10}$$

และ

$$\hat{\beta}_1 = \frac{n\left(\sum_{i=1}^{n}x_i y_i\right)-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2} \tag{2.11}$$

หรือ $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{2.12}$$

โดยทั่วไปตัวประมาณของวิธีภาวะน่าจะเป็นสูงสุดเป็นตัวประมาณที่มีคุณสมบัติทางสถิติ (Statistical Properties) ดีกว่าตัวประมาณที่ได้จากวิธีกำลังสองน้อยสุด ซึ่งตัวประมาณของวิธีภาวะน่าจะเป็นสูงสุดเป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียง (Unbiased) มีความแปรปรวนต่ำสุด (Minimum Variance) และมีความคงเส้นคงวา (Consistent) รวมไปถึงสถิติพอเพียง (Sufficient Statistics) [Douglas and Geoffrey, 2012]

2.4.3 คุณสมบัติของตัวประมาณค่าพารามิเตอร์ของการวิเคราะห์ถดถอยเชิงเส้นอย่างง่าย

ในทางทฤษฎีการอนุมานทางสถิติการตรวจสอบคุณสมบัติของตัวประมาณ มีคุณสมบัติที่สำคัญของตัวประมาณคือ ความไม่เอนเอียง (Unbiased) ซึ่งจะแสดงให้ทราบว่าโดยเฉลี่ยแล้วตัวประมาณนั้น ๆ ไม่มีแนวโน้มให้ค่าประมาณสูงหรือต่ำกว่าค่าพารามิเตอร์ และคุณสมบัติของตัวประมาณอีกข้อหนึ่งที่สำคัญคือ ค่าความแปรปรวนต่ำสุด (Minimum Variance) เป็นค่าที่แสดงให้ทราบว่าโดยเฉลี่ยแล้วกำลังสองของผลต่างระหว่างตัวประมาณกับค่าพารามิเตอร์เป็นเท่าใด

ตัวประมาณ $\hat{\beta}_0$ เป็นตัวประมาณที่ไม่เอนเอียงสำหรับ β_0 ซึ่งการพิจารณาคุณสมบัติความไม่เอนเอียงสามารถดำเนินการโดยหาค่าคาดหวังของตัวประมาณ $\hat{\beta}_0$ ซึ่งได้ทำการพิสูจน์ได้ดังนี้

$$\begin{aligned} E\left(\hat{\beta}_0\right) &= E\left(\bar{y}-\beta_1\bar{x}\right) \\ &= E\left(\bar{y}\right)-E\left(\beta_1\bar{x}\right) \\ &= E\left(\frac{\sum_{i=1}^{n}y_i}{n}\right)-\beta_1 E\left(\frac{\sum_{i=1}^{n}x_i}{n}\right) \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \frac{\beta_1}{n}\left(\sum_{i=1}^{n} x_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\beta_1\sum_{i=1}^{n} x_i - \frac{1}{n}\beta_1\sum_{i=1}^{n} x_i$$
$$= \beta_0$$

และตัวประมาณ $\hat{\beta}_1$ เป็นตัวประมาณที่ไม่เอนเอียงสำหรับ β_1 ซึ่งการพิจารณาคุณสมบัติความไม่เอนเอียงสามารถดำเนินการโดยหาค่าคาดหวังของตัวประมาณ $\hat{\beta}_1$ ซึ่งได้ทำการพิสูจน์ได้ดังนี้

จาก $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$E(\hat{\beta}_1) = E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$= E\left(\sum_{i=1}^{n} k_i y_i\right) \qquad \text{โดยที่ } k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sum_{i=1}^{n} k_i E(y_i)$$
$$= \sum_{i=1}^{n} k_i E(\beta_0 + \beta_1 x_i)$$
$$= \beta_0\sum_{i=1}^{n} k_i + \beta_1\sum_{i=1}^{n} k_i x_i \qquad ; \sum_{i=1}^{n} k_i = 0 \text{ และ } \sum_{i=1}^{n} k_i^2 = 1$$
$$= \beta_1$$

และคุณสมบัติของตัวประมาณที่สำคัญอีกประการหนึ่งคือ ความแปรปรวนต่ำสุด ตัวประมาณ $\hat{\beta}_0$ และตัวประมาณ $\hat{\beta}_1$ เป็นตัวประมาณที่มีความแปรปรวนต่ำสุด ซึ่งหาความแปรปรวนได้ดังนี้

$$Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

และ $$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

[พิษณุ เจียวคุณ, 2548]

2.4.4 การประเมินความผิดพลาดและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่าย

การวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเป็นการวิเคราะห์ความสัมพันธ์ระว่างตัวแปรอิสระหนึ่งตัว และตัวแปรตามหนึ่งตัว ซึ่งความสัมพันธ์ระหว่าง 2 ตัวแปรนี้เป็นไปในลักษณะความสัมพันธ์เชิงเส้นตรง การประเมินความผิดพลาดและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่าย จึงต้องทำการพิจารณาจากค่าความสัมพันธ์ของตัวแปรอิสระ และตัวแปรตาม ซึ่งความสัมพันธ์ต้องเป็นไปในลักษณะเชิงเส้น นั่นคือค่าที่ใช้ในการพิจารณาความสัมพันธ์คือ ค่าสัมประสิทธ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient : R) ซึ่งหาได้ดังนี้

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$ เมื่อ n คือขนาดตัวอย่าง

ซึ่ง $-1\le R\le 1$

นอกจากพิจารณาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามแล้วนั้น การนำการวิเคราะห์ถดถอยเชิงเส้นไปใช้ต้องคำนึงถึงข้อสมมติเบื้องต้นของการวิเคราะห์ด้วย ดังที่ได้กล่าวมาแล้วข้างต้นในหัวข้อที่ 2.4.1 และเมื่อทำการวิเคราะห์เพื่อหาตัวแบบถดถอยสำหรับนำไปใช้ในการพยากรณ์หรือคาดการณ์นั้น ต้องตรวจสอบความถูกต้องของตัวแบบถดถอยถึงความแม่นยำในการพยากรณ์ ซึ่งสามารถพิจารณาได้จากค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination : R^2) ซึ่ง $0\le R^2\le 1$

และค่าที่ต้องนำมาพิจารณาร่วมกันกับค่าสัมประสิทธ์สหสัมพันธ์ และสัมประสิทธ์การตัดสินใจคือค่าคลาดเคลื่อนเฉลี่ยกำลังสอง (Mean Square Error: MSE) เป็นค่าที่บอกให้ทราบว่าโดยเฉลี่ยแล้วค่าจริงต่างจากค่าที่ได้จากการพยากรณ์มากน้อยเพียงใด ซึ่งหาได้ดังนี้

$$MSE=\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-2}$$

[Douglas and Geoffrey, 2012]

2.5 การวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างถูกสุ่มตัวอย่างแบบชุดลำดับดับได้ดุล

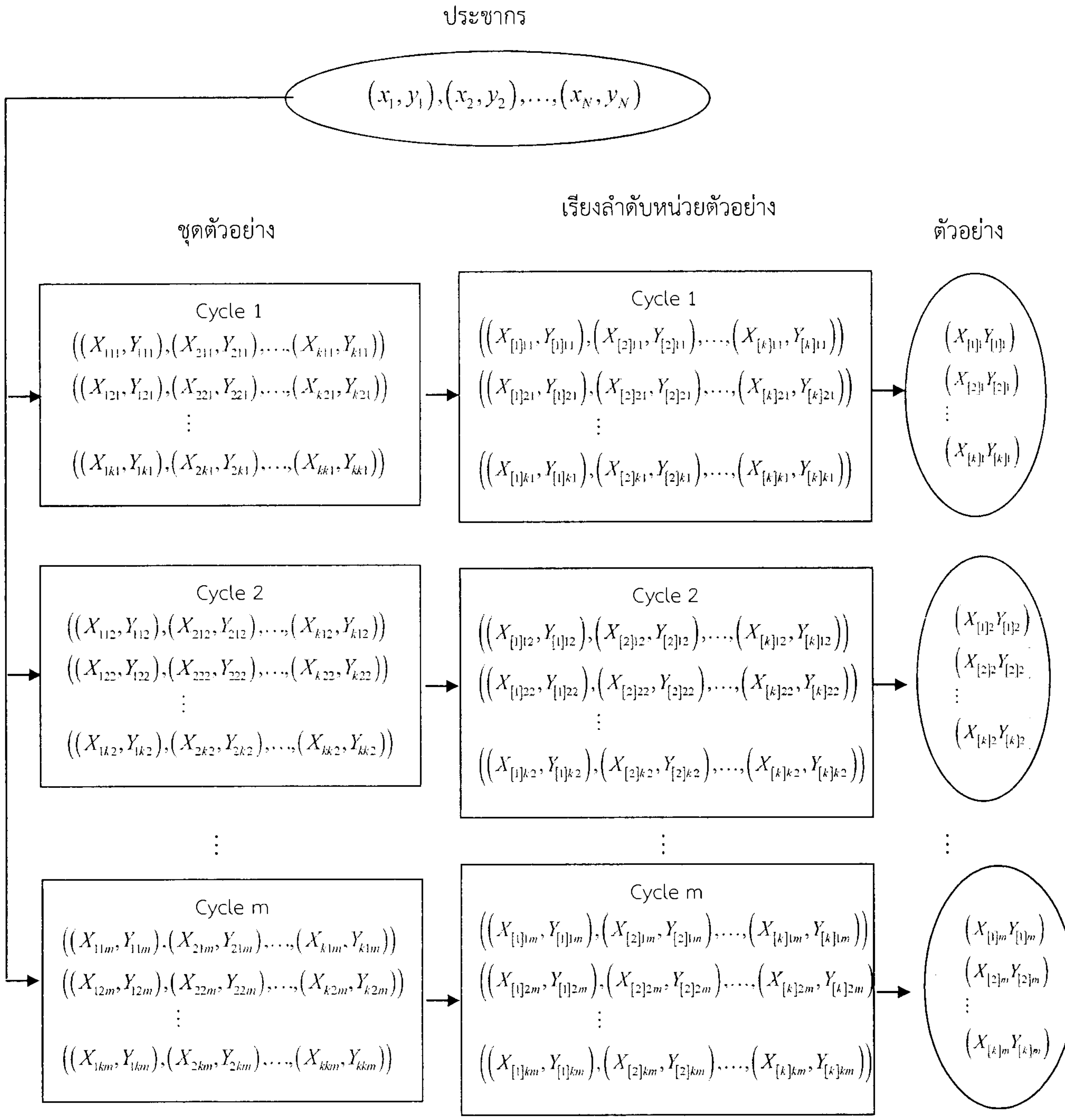
2.5.1 การสุ่มตัวอย่างสำหรับตัวแปรอิสระเพื่อทำการวัดค่าของตัวแปรตาม

การวิเคราะห์ถดถอยเชิงเส้นอย่างง่าย เป็นการศึกษาถึงความสัมพันธ์ของตัวแปรอิสระและตัวแปรตามของข้อมูลตัวอย่าง เพื่อทำการสร้างตัวแบบถดถอยเชิงเส้น $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_i$ จากตัวแบบจะเห็นว่า ถ้าหากตัวแปรอิสระ (X_i) สามารถอธิบายค่าของตัวแปรตาม (Y_i) ได้อย่างถูกต้องแม่นยำ หรือสามารถอธิบายสภาพความเป็นไปของตัวแปรตามได้ใกล้เคียงกับค่าจริงแล้วนั้น การวัดค่าตัวแปรตามด้วยค่าของตัวแปรอิสระก็จะสามารถนำมาใช้ในการแก้ปัญหาสำหรับการประมาณค่าตัวแปรตามที่มีข้อจำกัดไม่ว่าจะเป็นในด้านที่ว่าตัวแปรตามไม่สามารถวัดค่าได้โดยตรง หรือวัดค่าตัวแปรตามได้ยาก รวมทั้งข้อจำกัดเรื่องค่าใช้จ่าย เวลาในการศึกษา และบุคลากรซึ่งต้องใช้งบประมาณที่สูง รวมไปถึงข้อมูลที่อาจหาได้ยาก

ในปี 2001 Hani และ Faisal ได้ศึกษาการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายโดยใช้ตัวอย่างจากการสุ่มตัวอย่างแบบชุดลำดับ [Hani and Faisal, 2001] และในปี 2002 Chen ได้แนะนำตัวประมาณค่าเฉลี่ยประชากรของตัวแบบถดถอยสำหรับการสุ่มตัวอย่างแบบชุดลำดับ ซึ่งได้ทำการเรียงลำดับหน่วยตัวอย่างโดยตัว

แปรอิสระ จากนั้นทำการวัดค่าตัวแปรอิสระ และตัวแปรตามสำหรับการประมาณค่าเฉลี่ยประชากร [Chen, 2002]

สำหรับการวิจัยนี้จะทำการสุ่มตัวอย่างจากประชากรที่มีความสัมพันธ์กัน หรือมีข้อมูลประชากรสำหรับตัวแปรตาม (Y) และตัวแปรอิสระ (X) ซึ่งมีความสัมพันธ์กันเชิงเส้นตรง จากนั้นจะทำการสุ่มตัวอย่างจากข้อมูลประชากรทั้งตัวแปรตาม และตัวแปรอิสระด้วยการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย ดังขั้นตอนการสุ่มตัวอย่างสำหรับสร้างตัวแบบถดถอยในภาพที่ 2.5



ภาพที่ 2.5 แสดงขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับตัวแบบถดถอยเชิงเส้นอย่างง่าย

จากแผนภาพที่ 2.5 แสดงให้เห็นถึงการสุ่มหน่วยประชากรที่เป็นตัวแปรอิสระ X_{rij} ซึ่ง X_{rij} คือหน่วยประชากรที่เป็นตัวแปรอิสระที่ถูกสุ่มมาเป็นตัวอย่างหน่วยที่ r ; r = 1, 2, 3, ..., k ชุดตัวอย่างที่ i ; i = 1, 2, 3, ..., k และรอบที่ j ; j = 1, 2, 3, ..., m และ Y_{rij} คือหน่วยประชากรที่เป็นตัวแปรตามที่ถูกสุ่มมาเป็นตัวอย่าง หน่วยที่ r ; r = 1, 2, 3, ..., k ชุดตัวอย่างที่ i ; i = 1, 2, 3, ..., k และรอบที่ j ; j = 1, 2, 3, ..., m

จากนั้นหน่วยตัวอย่างที่ถูกสุ่มมาในแต่ละชุดตัวอย่างจะถูกเรียงลำดับทั้งตัวแปรอิสระและตัวแปรตามจากน้อยไปหามาก นั่นคือ จะได้ $X_{[r]ij}$ เป็นตัวอย่างหน่วยในลำดับที่ r ; r = 1, 2, 3, ..., k ของชุดตัวอย่างที่ i ; i = 1, 2, 3, ..., k และรอบที่ j ; j = 1, 2, 3, ..., m และ $Y_{[r]ij}$ คือหน่วยตัวอย่างที่เป็นตัวแปรตามในลำดับที่ r ; r = 1, 2, 3, ..., k ของชุดตัวอย่าง ที่ i ; i = 1, 2, 3, ..., k และรอบที่ j ; j = 1, 2, 3, ..., m

หน่วยตัวอย่างของชุดตัวอย่างที่ 1 ในรอบที่ 1 ที่อยู่ในลำดับที่ 1 จะถูกเลือกเป็นตัวอย่าง นั่นคือ $X_{[1]1}Y_{[1]1}$ จากนั้นหน่วยตัวอย่างในชุดตัวอย่างที่ 2 ของรอบที่ 1 ที่อยู่ในลำดับที่ 2 จะถูกเลือกเป็นตัวอย่าง นั่นคือ $X_{[2]1}Y_{[2]1}$ ทำแบบนี้ไปเรื่อย ๆ จนกระทั่งถึงชุดตัวอย่างที่ k รอบที่ m

ดังนั้นจากสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับการวิเคราะห์ถดถอยเชิงเส้นอย่างง่าย จะได้ตัวอย่างขนาด mk หน่วย จากประชากรขนาด N หน่วย คือ

$$\left(X_{[1]1}Y_{[1]1}\right),\left(X_{[2]1}Y_{[2]1}\right),\ldots,\left(X_{[k]1}Y_{[k]1}\right),\ldots,\left(X_{[1]m}Y_{[1]m}\right),\left(X_{[2]m}Y_{[2]m}\right),\ldots,\left(X_{[k]m}Y_{[k]m}\right)$$

2.5.2 การประมาณค่าพารามิเตอร์สำหรับการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างถูกสุ่มแบบชุดลำดับดับได้ดุล

ในการสุ่มตัวอย่างแบบชุดลำดับได้ดุลตัวประมาณค่าคุณลักษณะประชากรของตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เช่นตัวประมาณค่าเฉลี่ยประชากร พบว่ามีประสิทธิภาพดีกว่าตัวประมาณค่าคุณลักษณะที่ได้จากการสุ่มตัวอย่างอย่างง่าย ซึ่งตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลมา m รอบ รอบละ k ชุดตัวอย่าง สำหรับข้อมูลประชากรที่ถูกสุ่มมาเป็นคู่ คือ

$$\begin{array}{c}
\left(X_{[1]1}Y_{[1]1}\right),\left(X_{[2]1}Y_{[2]1}\right),\ldots,\left(X_{[k]1}Y_{[k]1}\right)\\
\left(X_{[1]2}Y_{[1]2}\right),\left(X_{[2]2}Y_{[2]2}\right),\ldots,\left(X_{[k]2}Y_{[k]2}\right)\\
\vdots \qquad\qquad\qquad \vdots\\
\left(X_{[1]m}Y_{[1]m}\right),\left(X_{[2]m}Y_{[2]m}\right),\ldots,\left(X_{[k]m}Y_{[k]m}\right)
\end{array}$$

จากตัวอย่างที่ได้ซึ่งมีขนาด $n = mk$ หน่วย เพื่อนำไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ซึ่งในการศึกษานี้จะได้ทำการสร้างตัวแบบพยากรณ์สำหรับประชากรที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งตัวแบบประชากรที่ศึกษาคือ

$$Y_{[r]iBRSS} = \beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} + \varepsilon_{[r]i} \qquad ;r = 1,2,3,\ldots,k \quad i = 1,2,3,\ldots,m$$

เมื่อ β_{0BRSS} และ β_{1BRSS} คือ พารามิเตอร์สำหรับการสุ่มแบบชุดลำดับได้ดุล

$\varepsilon_{[r]i}$ คือ ความคลาดเคลื่อนของลำดับที่ r รอบที่ i

$Y_{[r]iBRSS}$ คือ ตัวแปรตามสำหรับการสุ่มแบบชุดลำดับได้ดุล

$X_{[r]iBRSS}$ คือ ตัวแปรอิสระสำหรับการสุ่มแบบชุดลำดับได้ดุล

จากข้อสมมติเบื้องต้นเกี่ยวกับตัวแบบถดถอยเชิงเส้นอย่างง่าย ซึ่งในที่นี้ $Y_{[r]iBRSS}$ เป็นตัวแปรสุ่มที่มีการแจกแจงความน่าจะเป็น $f\left(Y_{[r]i} \mid X_{[r]i}\right)$ ค่าเฉลี่ยหรือค่าคาดหวังของ $Y_{[r]iBRSS}$ จะอยู่บนเส้นตรงซึ่งเป็นเส้นถดถอยของประชากร ซึ่งหาได้ดังนี้

$$\begin{aligned} E\left(Y_{[r]iBRSS}\right) &= E\left(\beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} + \varepsilon_{[r]i}\right) \\ &= \beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} \end{aligned}$$

และค่าความคลาดเคลื่อน $\varepsilon_{[r]i}$ มีการแจกแจงปกติ มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ σ^2 นั่นคือ $\varepsilon_{[r]i} \sim N\left(0,\sigma^2\right)$ เช่นเดียวกับการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายซึ่งตัวอย่างได้มาจากการสุ่มตัวอย่างแบบง่าย

ตัวแบบของตัวอย่างสำหรับพยากรณ์ค่าเฉลี่ยประชากร $\left(\hat{Y}_{[r]iBRSS}\right)$ คือ

$$\hat{Y}_{[r]iBRSS} = \hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}X_{[r]iBRSS}$$

เมื่อ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ คือ ตัวประมาณค่าพารามิเตอร์สำหรับการสุ่มแบบชุดลำดับได้ดุล

$\hat{Y}_{[r]iBRSS}$ คือ ตัวประมาณตัวแปรตามสำหรับการสุ่มแบบชุดลำดับได้ดุล

$X_{[r]iBRSS}$ คือ ตัวแปรอิสระสำหรับการสุ่มแบบชุดลำดับได้ดุล

สำหรับการประมาณค่าพารามิเตอร์ β_{0BRSS} และ β_{1BRSS} จะใช้วิธีกำลังสองน้อยที่สุด (Least – Squares Estimates) ประมาณค่า ซึ่งวิธีนี้มีหลักการประมาณค่าพารามิเตอร์คือ การประมาณค่าพารามิเตอร์ให้ค่าผลรวมกำลังสองของส่วนเหลือหรือความคลาดเคลื่อน (Residual) ที่น้อยที่สุด ซึ่งทำได้ดังนี้

กำหนดให้ $e_{[r]i}$ เป็นความคลาดเคลื่อนของตัวอย่างชุดที่ r รอบที่ i ซึ่ง $r = 1,2,3,\ldots,k$ และ $i = 1,2,\ldots,m$ จะได้ว่า

$$\begin{aligned} e_{[r]i} &= Y_{[r]iBRSS} - \hat{Y}_{[r]iBRSS} \\ &= Y_{[r]iBRSS} - \left(\hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}X_{[r]iBRSS}\right) \end{aligned}$$

ผลรวมกำลังสองของความคลาดเคลื่อนน้อยสุด คือ

$$\sum_{i=1}^{m}\sum_{r=1}^{k} e_{[r]i}^2 = \sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS} - \left(\hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}X_{[r]iBRSS}\right)\right]^2$$

กำหนดให้ Q แทน $\sum_{i=1}^{m}\sum_{r=1}^{k} e_{[r]i}^2$ จะได้ว่า

$$Q = \sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS} - \left(\hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}X_{[r]iBRSS}\right)\right]^2$$

เพื่อหาค่าที่น้อยที่สุดของ Q จึงหาอนุพันธ์ย่อยเทียบกับ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ ดังนี้

$$\frac{\partial Q}{\partial\hat{\beta}_{0BRSS}}=\frac{\partial}{\partial\hat{\beta}_{0BRSS}}\left[\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\left(\hat{\beta}_{0BRSS}+\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right)\right]^{2}\right]$$
$$=-2\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\hat{\beta}_{0BRSS}-\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right] \qquad (2.1)$$
$$\frac{\partial Q}{\partial\hat{\beta}_{1BRSS}}=\frac{\partial}{\partial\hat{\beta}_{1BRSS}}\left[\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\left(\hat{\beta}_{0BRSS}+\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right)\right]^{2}\right]$$
$$=-2\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\hat{\beta}_{0BRSS}-\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right]X_{[r]iBRSS} \qquad (2.2)$$

ให้สมการที่ (2.1) และ (2.2) เท่ากับ 0 จะได้

$$-2\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\hat{\beta}_{0BRSS}-\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right]=0 \qquad (2.3)$$
$$-2\sum_{i=1}^{m}\sum_{r=1}^{k}\left[Y_{[r]iBRSS}-\hat{\beta}_{0BRSS}-\hat{\beta}_{1BRSS}X_{[r]iBRSS}\right]X_{[r]iBRSS}=0 \qquad (2.4)$$

จากสมการที่ (2.3) และ (2.4) จะได้

$$\sum_{i=1}^{m}\sum_{r=1}^{k}Y_{[r]iBRSS}=mk\hat{\beta}_{0BRSS}+\hat{\beta}_{1BRSS}\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS} \qquad (2.5)$$
$$\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}Y_{[r]iBRSS}=\hat{\beta}_{0BRSS}\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}+\hat{\beta}_{1BRSS}\sum_{i=1}^{m}\sum_{r=1}^{k}X^{2}_{[r]iBRSS} \qquad (2.6)$$

แก้ระบบสมการที่ (2.5) และ (2.6) ได้ตัวประมาณสำหรับพารามิเตอร์ β_{0BRSS} และ β_{1BRSS} ดังนี้

$$\hat{\beta}_{1BRSS}=\frac{mk\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}Y_{[r]iBRSS}-\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}\sum_{i=1}^{m}\sum_{r=1}^{k}Y_{[r]iBRSS}}{mk\sum_{i=1}^{m}\sum_{r=1}^{k}X^{2}{}_{[r]iBRSS}-\left(\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}\right)^{2}}$$

และ

$$\hat{\beta}_{0BRSS}=\frac{\sum_{i=1}^{m}\sum_{r=1}^{k}Y_{[r]iBRSS}}{mk}-\hat{\beta}_{1BRSS}\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}}{mk}\right)$$

2.5.3 คุณสมบัติของตัวประมาณค่าพารามิเตอร์ของการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างถูกสุ่มตัวอย่างแบบชุดลำดับดับได้ดุล

การตรวจสอบคุณสมบัติของตัวประมาณ ในทางทฤษฎีการอนุมานทางสถิติ คุณสมบัติของตัวประมาณที่สำคัญคือ ความไม่เอนเอียง และความแปรปรวนต่ำสุด คุณสมบัติความไม่เอนเอียงสามารถดำเนินการโดยหาค่าคาดหวังของตัวประมาณได้ดังนี้

สำหรับตัวประมาณพารามิเตอร์ $\hat{\beta}_{0BRSS}$ ดำเนินการหาค่าคาดหวังดังต่อไปนี้

$$E\left(\hat{\beta}_{0BRSS}\right)=E\left[\frac{\sum_{i=1}^{m}\sum_{r=1}^{k}Y_{[r]iBRSS}}{mk}-\hat{\beta}_{1BRSS}\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k}X_{[r]iBRSS}}{mk}\right)\right]$$

$$= E\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k} Y_{[r]iBRSS}}{mk}\right) - \hat{\beta}_{1BRSS} E\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}}{mk}\right)$$

$$= \frac{1}{mk}\left(\sum_{i=1}^{m}\sum_{r=1}^{k}\left(\beta_{0BRSS} + \beta_{1BRSS} X_{[r]iBRSS}\right)\right) - \frac{\hat{\beta}_{1BRSS}}{mk} E\left(\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}\right)$$

$$= \beta_{0BRSS}$$

จากข้างต้นจะได้ว่าตัวประมาณ $\hat{\beta}_{0BRSS}$ เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียงและตัวประมาณพารามิเตอร์ $\hat{\beta}_{1BRSS}$ เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียงเช่นกัน ซึ่งสามารถดำเนินการหาค่าคาดหวังเช่นเดียวกัน

2.6 งานวิจัยที่เกี่ยวข้อง

McIntyre ได้เสนอวิธีการสุ่มตัวอย่างซึ่งวิธีการสุ่มตัวอย่างมีดังนี้ ขั้นแรกสุ่มตัวอย่างขนาด n หน่วยขึ้นมา n ชุดตัวอย่าง จากนั้นทำการเรียงลำดับหน่วยตัวอย่างแต่ละชุดตัวอย่างด้วยการคาดคะเน และจะเลือกหน่วยตัวอย่างที่อยู่ในลำดับที่ 1 จาก n หน่วย ในชุดตัวอย่างที่ 1 เป็นตัวอย่าง และเลือกตัวอย่างที่อยู่ในลำดับ 2 ในชุดตัวอย่างที่ 2 และสุดท้ายเลือกตัวอย่างที่อยู่ในลำดับที่ n จากชุดตัวอย่างที่ n เป็นตัวอย่าง ดังนั้นจะได้ตัวอย่างทั้งหมด n หน่วย สำหรับวิธีการสุ่มตัวอย่างนี้ ตัวอย่างที่ได้จะนำมาประมาณค่าลักษณะประชากร ซึ่งการประยุกต์ของวิธีการเลือกตัวอย่างแบบลำดับนี้มีวัตถุประสงค์เพื่อประมาณหญ้าในทุ่งหญ้า [McIntyre, 1952]

Kowalczyk ได้ศึกษาการสุ่มตัวอย่างแบบชุดลำดับได้ดุลจากประชากรอันตะซึ่งได้ให้ข้อตกลงเบื้องต้นเกี่ยวกับการสุ่มตัวอย่างแบบชุดลำดับสำหรับประชากรอันตะดังนี้ ในการสุ่มตัวอย่างแบบชุดลำดับภายใต้ประชากรอันตะขนาด N หน่วย และได้กำหนดให้ Y_j เป็นประชากรหน่วยที่ j ซึ่ง $j = 1,2,\ldots,N$ และค่าของ Y_j เป็นค่าคงที่ซึ่งติดอยู่กับหน่วยประชากร Y_j แต่ละหน่วย และได้ค่า ตัวประมาณเฉลี่ยประชากร คือ $\bar{y}_{RSS} = \frac{1}{n}\sum_{i=1}^{m}\sum_{r=i}^{k} y_{[r]i}$ ซึ่งมีคุณสมบัติของตัวประมาณ ดังนี้

1) $\bar{y}_{RSS}$ ภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลจากประชากรอันตะเป็นตัวประมาณที่ไม่เอนเอียงของ $\bar{Y}$

2) ความแปรปรวนของ $\bar{y}_{RSS}$ มีดังนี้

$$Var(\bar{y}_{RSS}) = \frac{1}{n}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})] - \bar{Y}\right)^2\right\} \text{ เมื่อ } S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(Y_j - \bar{Y}\right)^2$$

3) $\bar{y}_{RSS}$ ภายใต้การสุ่มตัวอย่างแบบชุดลำดับมีประสิทธิภาพมากกว่า (Efficient) $\bar{y}_{SRS}$ ภายใต้การสุ่มตัวอย่างแบบง่ายแบบใส่คืน

4) $y_{RSS} = N\bar{y}_{RSS}$ เป็นตัวประมาณที่ไม่เอนเอียงของยอดรวมประชากร Y

5) ความแปรปรวนของตัวประมาณ y_{RSS} ของยอดรวมประชากร มีดังนี้

$$Var(y_{RSS}) = \frac{N^2}{n}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})] - \bar{Y}\right)^2\right\}$$

6) ตัวประมาณ y_{RSS} มีประสิทธิภาพมากกว่า ตัวประมาณ $y_{SRS} = N\bar{y}_{SRS}$ ภายใต้การสุ่มตัว-อย่างแบบง่ายใส่คืน [Kowalczyk, 2004]

Hani และ Faisal ได้ศึกษาการวิเคราะห์ถดถอยโดยใช้ตัวอย่างแบบชุดลำดับ (On Regression Analysis Using Ranked Set Sampling) ซึ่งได้ทำการสุ่มตัวอย่างด้วยตัวแปรอิสระแล้วทำการวัดค่าตัวแปรตาม โดยมีตัวแบบคือ $Y_{(j)k} = \alpha + \beta X_{(j)k} + \varepsilon_{jk}$ เมื่อ $j = 1,2,\ldots,r$ และ $k = 1,2,\ldots,m$ ตัวประมาณพารามิเตอร์ที่ได้คือ

$\hat{\alpha} = \bar{Y}_{RSS} - \hat{\beta}_{RSS}\bar{X}_{RSS}$ และ

$$\hat{\beta}_{RSS} = \frac{\sum_{k=1}^{m}\sum_{i=1}^{r}\left(X_{(j)k} - \bar{X}_{RSS}\right)Y_{[j]k}}{\sum_{k=1}^{m}\sum_{i=1}^{r}\left(X_{(j)k} - \bar{X}_{RSS}\right)^2}$$

[Hani and Faisal, 2001]

Arzu และ Esin ได้ศึกษาการประมาณค่าพารามิเตอร์สำหรับตัวแบบถดถอยพหุคูณโดย (Multiple Linear Regression Models) เมื่อใช้วิธีการสุ่มตัวอย่างแบบชุดลำดับ โดยทำการศึกษาเปรียบเทียบวิธีการสุ่มตัวอย่างแบบ RSS และ SRS ในการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยพหุคูณ ซึ่งได้จำลองข้อมูลโดย Monte Carlo พบว่า ตัวประมาณพารามิเตอร์ซึ่งได้จากการสุ่มตัวอย่างแบบ RSS มีประสิทธิภาพมากกว่า ตัวประมาณที่สุ่มตัวอย่างมาจากวิธี SRS เมื่อมีขนาดตัวอย่างน้อย โดยศึกษากรณีที่ตัวแปรอิสระมี 2 ตัวแปร [Arzu and Esin, 2007]

บทที่ 3
การดำเนินการวิจัย

3.1 ความนำ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล เกณฑ์ที่ใช้ในการตรวจสอบความถูกต้องและความเชื่อถือได้พิจารณาจากค่าสัมประสิทธิ์การตัดสินใจ และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ซึ่งมีรายละเอียดเกี่ยวกับแผนการวิจัย ขั้นตอนการดำเนินการวิจัย และขั้นตอนการจำลองข้อมูลที่ใช้ในการวิจัยนำเสนอตามลำดับดังต่อไปนี้

3.2 แผนการวิจัย

ตามวัตถุประสงค์การศึกษาที่ได้กล่าวมาแล้วในบทที่ 1 และได้ศึกษาตัวแบบถดถอยอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งได้หาตัวประมาณพารามิเตอร์พร้อมทั้งตรวจสอบคุณสมบัติของตัวประมาณค่าพารามิเตอร์ดังในบทที่ 2 ผู้วิจัยจึงได้ทำการกำหนดสถานการณ์เพื่อทำการตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุลโดยการจำลองข้อมูล ดังต่อไปนี้

1. ประชากรที่ศึกษาสร้างขึ้นมาภายใต้การแจกแจงปกติ (Normal Distribution) โดยกำหนดให้ข้อมูลประชากรมีความสัมพันธ์กัน นั่นคือข้อมูลประชากรตัวแปรอิสระ (x) และตัวแปรตาม (y) มีความสัมพันธ์กัน

2. กำหนดค่าสหสัมพันธ์ ρ (Correlations) สำหรับตัวแปรอิสระและตัวแปรตาม แบ่งออกเป็น 3 ระดับความสัมพันธ์ ได้แก่

    1) ระดับความสัมพันธ์มาก คือ 0.71-1.00

    2) ระดับความสัมพันธ์ปานกลาง คือ 0.31-0.70

    3) ระดับความสัมพันธ์ต่ำ คือ 0.00-0.30 [Ratner, 2009]

3. ขนาดประชากร (N) ที่ศึกษาเท่ากับ 1,000

4. ขนาดของตัวอย่าง (n) ที่ศึกษา คือ $n=5$, 10, 30, 40 และ 50 [Voorhis and Morgan, 2007] โดยที่ $n = mk$ ซึ่ง k คือ ขนาดตัวอย่างในการสุ่มแต่ละชุดตัวอย่างและเป็นจำนวนชุดตัวอย่างของการสุ่มแต่ละรอบ และ m คือ จำนวนรอบในการสุ่มตัวอย่าง

5. ทำการทดลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ที่กำหนด (จำนวนการทดลอง 1,000 ครั้ง ในแต่ละสถานการณ์ เป็นจำนวนการทดลองที่มากพอในการจำลองข้อมูล) [Levis, 2017]
[Sokolowsak and Banks, 2009]

จากแผนการจำลองข้อมูล จะทำการจำลองข้อมูลประชากรแล้วทำการสุ่มตัวอย่างแบบชุดลำดับได้ดุล เพื่อทำการสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย เพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่าย

3.3 ขั้นตอนการดำเนินการวิจัย

ขั้นตอนการดำเนินการวิจัยมีดังต่อไปนี้

1. จำลองหน่วยประชากร กำหนดขนาดของประชากร และกำหนดค่าสหสัมพันธ์ ρ สำหรับตัวแปรอิสระและตัวแปรตาม แบ่งออกเป็น 3 ระดับความสัมพันธ์ ตามแผนการวิจัย

2. สุ่มตัวอย่างแบบชุดลำดับได้ดุลจากประชากรที่ถูกจำลองขึ้นมาในข้อ 1. ซึ่งมีขนาดตัวอย่างตามแผนการจำลองข้อมูล ภายใต้การแจกแจงปกติ ซึ่งมีกรณีที่ทำการศึกษาทั้งหมด 23 กรณี ดังต่อไปนี้

ตารางที่ 3.1 แสดงกรณีของขนาดประชากร และขนาดตัวอย่างที่ทำการวิจัย

ขนาดประชากร (N)	**ขนาดตัวอย่าง** $(n = mk)$	**ขนาดประชากร** (N)	**ขนาดตัวอย่าง** $(n = mk)$
N = 1,000	n = 5 m = 1, k = 5 n = 10 m = 1, k = 10 m = 2, k = 5 m = 5, k = 2 n = 30 m = 1, k = 30 m = 2, k = 15 m = 3, k = 10 m = 5, k = 6 m = 6, k = 5 m = 10, k = 3 m = 15, k = 2	N = 1,000	n = 40 m = 1, k = 40 m = 2, k = 20 m = 4, k = 2 m = 5, k = 8 m = 8, k = 5 m =10, k = 5 m = 20, k = 2 n = 50 m = 1, k = 50 m = 2, k = 25 m = 5, k = 10 m = 10, k = 5 m = 25, k = 2

3. สร้างตัวแบบถดถอยอย่างง่ายจากตัวอย่างที่ถูกสุ่มด้วยการสุ่มตัวอย่างแบบชุดลำดับได้ดุลโดยนำข้อมูลที่ได้ในข้อ 2. มาคำนวณเพื่อสร้างตัวแบบดังบทที่ 2 ดังนี้

ตัวแบบของตัวอย่างสำหรับพยากรณ์ค่าเฉลี่ยประชากร $\left(\hat{\bar{Y}}_{iBRSS}\right)$ คือ

$$\hat{\bar{Y}}_{iBRSS} = \hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}\bar{X}_{iBRSS}$$

ตัวประมาณสำหรับพารามิเตอร์ β_{0BRSS} และ β_{1BRSS} มีดังนี้

$$\hat{\beta}_{1BRSS} = \frac{m\sum_{i=1}^{m}\bar{X}_{iBRSS}\bar{Y}_{iBRSS} - \sum_{i=1}^{m}\bar{X}_{iBRSS}\sum_{i=1}^{m}\bar{Y}_{iBRSS}}{m\sum_{i=1}^{m}\bar{X}_{iBRSS}^{2} - \left(\sum_{i=1}^{m}\bar{X}_{iBRSS}\right)^{2}}$$

และ $$\hat{\beta}_{0BRSS} = \frac{\sum_{i=1}^{m} \overline{Y}_{iBRSS}}{m} - \hat{\beta}_{1BRSS} \left(\frac{\sum_{i=1}^{m} \overline{X}_{iBRSS}}{m} \right)$$

4. หาค่าสัมประสิทธิ์การตัดสินใจ (R^2) โดยหาได้จาก

$$R^2 = \left(\frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y_i\right)\right)^2}} \right)^2$$

5. หาค่าเฉลี่ยความคลาดเคลื่อนกำลังสองโดยหาได้จาก

$$MSE = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2}{n-2}$$

6. ดำเนินการตามข้อ 1 – 5 ซ้ำ จำนวน 1,000 ครั้งในแต่ละสถานการณ์

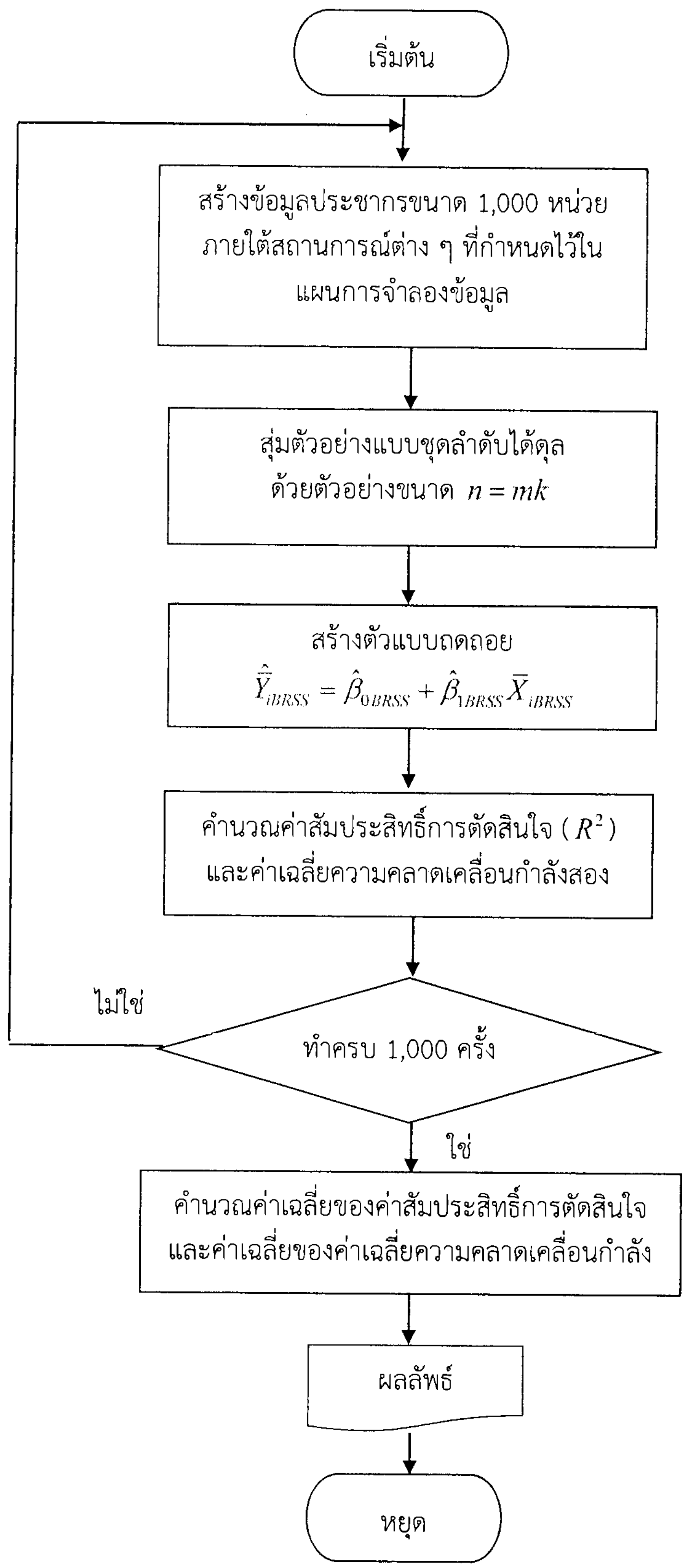
7. เปรียบเทียบค่าสัมประสิทธิ์การตัดสินใจ และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

8. สรุปผลการวิจัย

การสรุปผลการวิจัยเราจะทำการสรุปผลการวิจัยว่าที่ขนาดตัวอย่างใดและจำนวนรอบเท่าใดให้ค่าสัมประสิทธิ์การตัดสินใจสำหรับตัวแบบพยากรณ์ได้มากที่สุด และมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุด ที่ขนาดตัวอย่างและจำนวนรอบนั้น มีความเหมาะสมกับตัวแบบถดถอยมากที่สุดสำหรับแต่ละสถานการณ์ที่ได้กำหนดไว้ในแผนการวิจัย

3.4 ขั้นตอนการจำลองข้อมูล และการตรวจสอบความถูกต้องและความเชื่อถือได้ของตัวแบบ

ในการวิจัยครั้งนี้ผู้วิจัยได้ใช้โปรแกรม MINITAB 17.0 ซึ่งเป็นโปรแกรมสำเร็จรูปทางสถิติในการสร้างข้อมูล การสุ่มตัวอย่าง คำนวณค่าสัมประสิทธิ์การตัดสินใจ และคำนวณค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง โดยขั้นตอนในการจำลองข้อมูล และการตรวจสอบความถูกต้องและความเชื่อถือได้ของตัวแบบ แสดงดังภาพที่ 3.1



ภาพที่ 3.1 ขั้นตอนการจำลองข้อมูลและการตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล

บทที่ 4
ผลการศึกษา

4.1 ความนำ

ดังที่ได้ระบุไว้ในบทที่ 1 แล้วว่าการศึกษาในครั้งนี้ต้องการสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล และตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุล เพื่อเป็นแนวทาง ในการสร้างตัวแบบให้มีความถูกต้องเหมาะสม และความน่าเชื่อถือได้ภายใต้ทฤษฎีการสำรวจตัวอย่างและทฤษฎีการอนุมานทางสถิติ

ในการศึกษาตรวจสอบความถูกต้อง และความน่าเชื่อถือได้ของของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับดุลดำเนินการโดยทำการจำลองข้อมูลตามสถานการณ์ที่กำหนดไว้ในบทที่ 3 ซึ่งผลการศึกษาได้แบ่งออกเป็น 2 ส่วนดังต่อไปนี้

4.2 ผลการตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของของตัวแบบจากค่าสัมประสิทธิ์การตัดสินใจ (R^2) ในแต่ละสถานการณ์

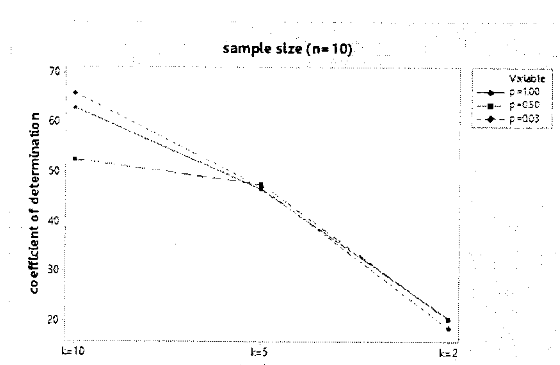
การวิจัยในส่วนนี้เป็นการตรวจสอบความถูกต้อง และความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อข้อมูลถูกสุ่มมาจากการสุ่มตัวอย่างแบบชุดลำดับดับได้ดุลจากค่าสัมประสิทธิ์การตัดสินใจ โดยได้ศึกษาข้อมูลประชากรขนาด 1,000 ภายใต้ประชากรที่มีการแจกแจงแบบปกติ ซึ่งได้กำหนดระดับความสัมพันธ์ของข้อมูลตัวแปรอิสระและตัวแปรตามไว้ 3 ระดับ คือข้อมูลมีความสัมพันธ์กันมาก ปานกลาง และน้อย ดังที่กล่าวในบทที่ 3 ซึ่งผลการวิจัยมีดังต่อไปนี้

ตารางที่ 4.1 ค่าสัมประสิทธิ์การตัดสินใจ (R^2) สำหรับกรณีต่าง ๆ จำแนกตามขนาดตัวอย่างและระดับความสัมพันธ์

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	R^2 (%)		
			$\rho = 1.00$	$\rho = 0.50$	$\rho = 0.03$
5	1	5	52.813	52.524	52.589
10	1	10	62.916	66.456	65.806
	2	5	46.477	47.433	46.818
	5	2	20.252	20.210	18.459
30	1	30	85.994	85.292	85.464
	2	15	77.163	74.969	74.286
	3	10	64.648	64.893	64.488
	5	6	49.445	48.696	49.815
	6	5	43.138	42.793	43.739
	10	3	25.104	25.910	25.259
	15	2	12.333	12.418	13.522

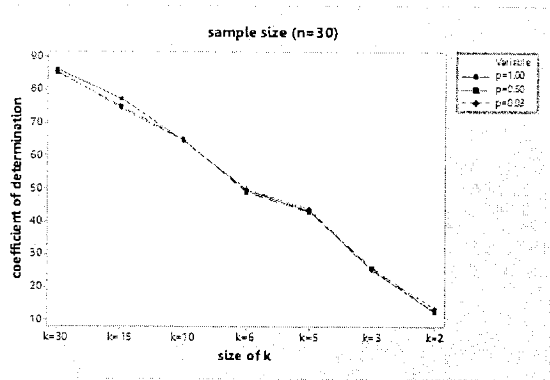
ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	R^2 (%)		
			$\rho = 1.00$	$\rho = 0.50$	$\rho = 0.03$
40	1	40	88.900	88.251	88.348
	2	20	79.376	78.702	79.122
	4	10	64.332	52.889	59.257
	5	8	57.726	57.483	58.236
	8	5	42.900	41.552	43.856
	10	4	34.425	35.194	34.120
	20	2	12.619	12.671	13.347
50	1	50	90.499	89.650	89.871
	2	25	82.707	81.446	82.453
	5	10	63.346	62.656	63.070
	10	5	41.178	41.100	42.550
	25	2	12.583	11.973	11.235

จากตารางที่ 4.1 พบว่าข้อมูลประชากรที่กำหนดค่าสหสัมพันธ์ (ρ) สำหรับตัวแปรอิสระและตัวแปรตามที่แตกต่างกันทั้ง 3 ระดับนั้น เมื่อนำตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย พบว่า เมื่อพิจารณาตามขนาดชุดตัวอย่าง(k) หรือจำนวนรอบ (m) ค่า R^2 ที่ได้จากตัวแบบถดถอยเชิงเส้นในแต่ละกรณีสำหรับตัวอย่างขนาด 5 10 30 40 และ 50 มีค่าใกล้เคียงกันที่ค่าสหสัมพันธ์ต่างกัน และที่ n = 5 มีค่าสัมประสิทธิ์การตัดสินใจที่ค่าสหสัมพันธ์ทั้ง 3 ระดับ ประมาณ 52% ส่วนที่ n =10, 30, 40 และ 50 ที่จำนวนการสุ่ม 1 รอบ ($m = 1$) มีค่าสัมประสิทธิ์การตัดสินใจที่ค่าสหสัมพันธ์ทั้ง 3 ระดับ สูงที่สุด คือประมาณ 65%, 85%, 88% และ 90% ตามลำดับ เมื่อนำผลการศึกษาที่ได้ในตารางที่ 4.1 มาสร้างกราฟแสดงได้ดังภาพที่ 4.1-4.4 สำหรับกรณีที่ n = 10, 30, 40 และ 50 ตามลำดับ ดังนี้



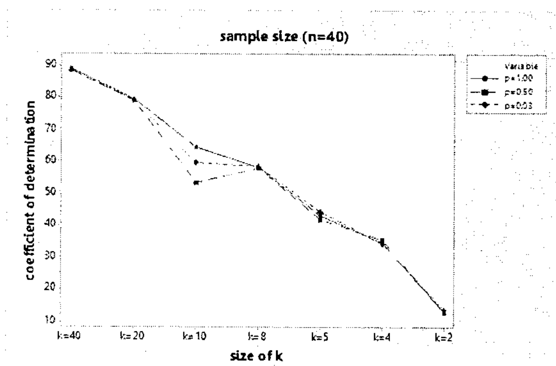
ภาพที่ 4.1 แสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 10

จากภาพที่ 4.1 ซึ่งแสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 10 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง(k) เท่ากับ 10, 5 และ 2 ตามลำดับ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) มีแนวโน้มลดลง เมื่อขนาดชุดตัวอย่าง(k) ลดลง



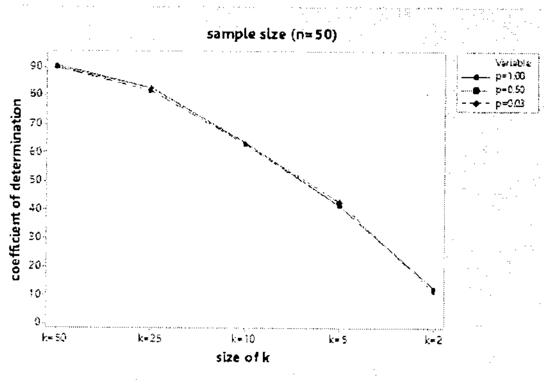
ภาพที่ 4.2 แสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 30

จากภาพที่ 4.2 ซึ่งแสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 30 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง(k) เท่ากับ 30, 15, 10, 5, 3 และ 2 ตามลำดับ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) มีแนวโน้มลดลง เมื่อขนาดชุดตัวอย่าง(k) ลดลง



ภาพที่ 4.3 แสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 40

จากภาพที่ 4.3 ซึ่งแสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 40 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง(k) เท่ากับ 40, 20, 10, 8, 4 และ 2 ตามลำดับ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) มีแนวโน้มลดลง เมื่อขนาดชุดตัวอย่าง(k) ลดลง



ภาพที่ 4.4 แสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 50

จากภาพที่ 4.4 ซึ่งแสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (R^2) กับขนาดชุดตัวอย่าง(k) ที่ขนาดตัวอย่าง 50 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง(k) เท่ากับ 50, 25, 10, 5 และ 2 ตามลำดับ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) มีแนวโน้มลดลง เมื่อขนาดชุดตัวอย่าง(k) ลดลง

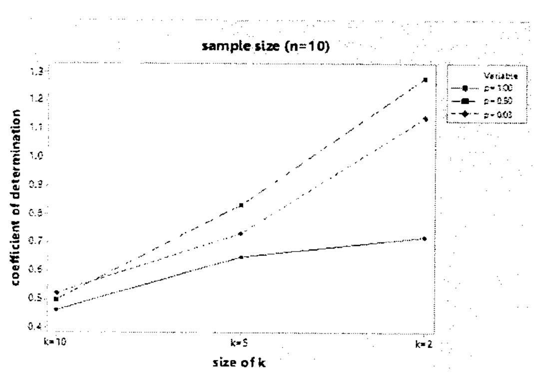
4.3 ผลการตรวจสอบถูกต้อง และความน่าเชื่อถือได้ของของตัวแบบจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสองในแต่ละสถานการณ์

ตารางที่ 4.2 ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) สำหรับกรณีต่าง ๆ ตามแผนการจำลองข้อมูล

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	MSE		
			$\rho = 1.00$	$\rho = 0.50$	$\rho = 0.03$
5	1	5	0.527	0.897	0.989
10	1	10	0.459	0.497	0.519
	2	5	0.649	0.833	0.732
	5	2	0.718	1.280	1.140
30	1	30	0.131	0.191	0.154
	2	15	0.198	0.308	0.307
	3	10	0.407	0.434	0.444
	5	6	0.523	0.702	0.607
	6	5	0.619	0.732	0.661
	10	3	0.751	1.060	0.885
	15	2	0.928	1.333	0.952

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	MSE		
			$\rho = 1.00$	$\rho = 0.50$	$\rho = 0.03$
40	1	40	0.102	0.177	0.076
	2	20	0.195	0.298	0.216
	4	10	0.314	0.251	0.215
	5	8	0.416	0.638	0.485
	8	5	0.541	0.753	0.742
	10	4	0.692	0.951	0.712
	20	2	0.948	1.224	1.002
50	1	50	0.093	0.120	0.111
	2	25	0.192	0.216	0.211
	5	10	0.370	0.460	0.424
	10	5	0.565	0.834	0.630
	25	2	0.824	1.278	1.070

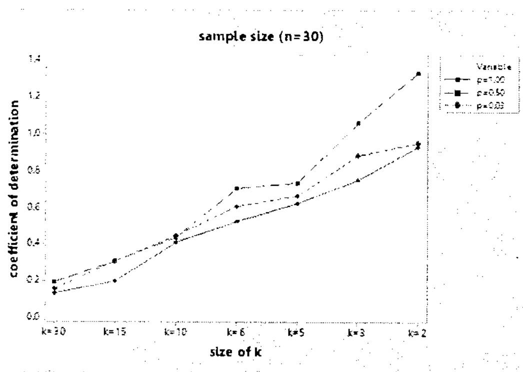
จากตารางที่ 4.2 พบว่าที่ข้อมูลประชากรที่กำหนดค่าสหสัมพันธ์ (ρ) สำหรับตัวแปรอิสระและตัวแปรตามที่แตกต่างกันทั้ง 3 ระดับนั้น เมื่อนำตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย พบว่า ที่ $n =$ 5 และ ρ =1.00 ตัวแบบที่ได้ให้ค่าความคลาดเคลื่อนกำลังสองน้อยที่สุด ส่วนที่ $n =$ 10, 30, 40, 50 ที่ $m =$ 1 ตัวแบบที่ได้ให้ค่าความคลาดเคลื่อนกำลังสองน้อยที่สุดสำหรับทุกค่าสหสัมพันธ์ (ρ) และจะมีค่าความคลาดเคลื่อนกำลังสองเพิ่มขึ้นเมื่อมีจำนวนรอบของการสุ่มตัวอย่าง (m)เพิ่มขึ้น หรือขนาดชุดตัวอย่าง (k) ลดลง เมื่อนำผลการศึกษาที่ได้ในตารางที่ 4.2 มาสร้างกราฟแสดงได้ดังภาพที่ 4.5-4.8 สำหรับกรณีที่ n = 10, 30, 40 และ 50



ภาพที่ 4.5 แสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 10

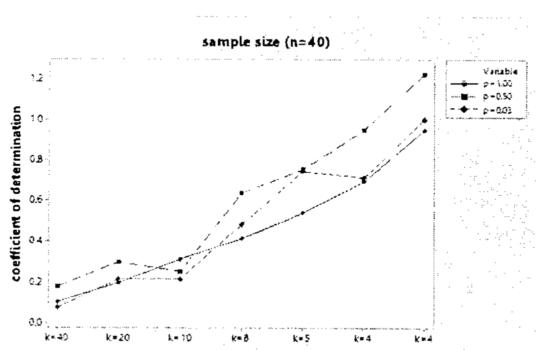
จากภาพที่ 4.5 ซึ่งแสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 10 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมี

ความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง (k) เท่ากับ 10, 5 และ 2 ตามลำดับ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) มีแนวโน้มเพิ่มขึ้น เมื่อขนาดชุดตัวอย่าง(k) ลดลง



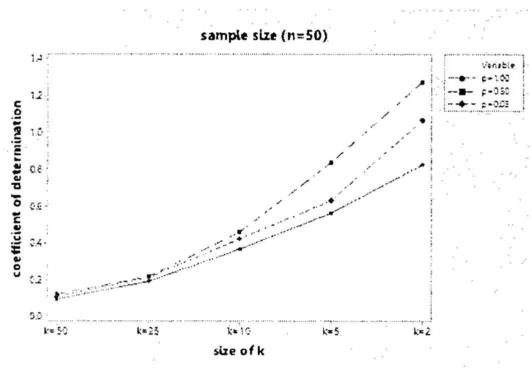
ภาพที่ 4.6 แสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 30

จากภาพที่ 4.6 ซึ่งแสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 30 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง (k) เท่ากับ 30, 15, 10, 6, 5, 3 และ 2 ตามลำดับ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) มีแนวโน้มเพิ่มขึ้น เมื่อขนาดชุดตัวอย่าง (k) ลดลง



ภาพที่ 4.7 แสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 40

จากภาพที่ 4.7 ซึ่งแสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 40 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง (k) เท่ากับ 40, 20, 10, 8, 5, 4 และ 2 ตามลำดับ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) มีแนวโน้มเพิ่มขึ้น เมื่อขนาดชุดตัวอย่าง (k) ลดลง



ภาพที่ 4.8 แสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 50

จากภาพที่ 4.8 ซึ่งแสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k) ที่ขนาดตัวอย่าง 50 ของข้อมูลตัวอย่างที่ถูกสุ่มมาจากกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันทั้ง 3 ระดับ คือ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ จากกราฟจะพบว่า ที่ขนาดชุดตัวอย่าง (k) เท่ากับ 50, 25, 10, 5 และ 2 ตามลำดับ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) มีแนวโน้มเพิ่มขึ้น เมื่อขนาดชุดตัวอย่าง (k) ลดลง

บทที่ 5
สรุปและอภิปรายผลการศึกษา

5.1 ความนำ

จากผลการวิจัยในบทที่ 4 ซึ่งได้แบ่งออกเป็น 2 ส่วน คือส่วนที่ 1 เป็นการตรวจสอบถูกต้อง และความน่าเชื่อถือได้ของของตัวแบบจากค่าสัมประสิทธิ์การตัดสินใจ (R^2) ในแต่ละสถานการณ์ และส่วนที่ 2 เป็นผลการตรวจสอบความถูกต้อง และความน่าเชื่อถือได้ของของตัวแบบจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) ในแต่ละสถานการณ์ ดังนั้นเพื่อหาข้อสรุปว่าตัวแบบถดถอยเชิงเส้นอย่างง่ายที่สร้างขึ้นจากตัวอย่างซึ่งถูกสุ่มด้วยวิธีการสุ่มตัวอย่างแบบชุดลำดับได้ดุลมีความเหมาะสมมากที่สุดกับสถานการณ์ที่ได้จำลองขึ้นมาในแต่ละสถานการณ์จึงนำผลการวิจัยทั้ง 2 ส่วนมาพิจารณาร่วมกัน ซึ่งเกณฑ์เบื้องต้นที่ใช้ในการพิจารณาคือสถานการณ์ใดให้ค่าสัมประสิทธิ์การตัดสินใจมากพอ และมีค่าความคลาดเคลื่อนกำลังสองน้อยที่สุดจะพิจารณาได้ว่าสถานการณ์นั้นเหมาะสมที่สุดสำหรับการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อนำไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย โดยผลการวิจัยสามารถสรุปได้ดังนี้

5.2 สรุปผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์คือ สร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งตัวแบบถดถอยคือ $Y_{[r]iBRSS} = \beta_{0BRSS} + \beta_{1BRSS} X_{[r]iBRSS} + \varepsilon_{[r]i}$ โดยมีตัวประมาณค่าพารามิเตอร์ คือ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งได้จำลองข้อมูลตามสถาณการณ์ต่าง ๆ พบว่าที่ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 สำหรับกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันที่ $\rho = 1.00$, $\rho = 0.50$ และ $\rho = 0.03$ ตัวแบบถดถอยที่ศึกษาจะให้ค่า R^2 สูงขึ้นเมื่อขนาดตัวอย่าง (n) เพิ่มมากขึ้น และตัวแบบถดถอยที่ยังให้ค่า R^2 สูงที่สุด เมื่อทำการสุ่มตัวอย่าง 1 รอบ หรือ $m = 1$ ซึ่งที่ขนาดตัวอย่าง n = 5 ได้ค่า R^2 ประมาณ 52% ขนาดตัวอย่าง n = 10 ได้ค่า R^2 ประมาณ 62% – 65% ขนาดตัวอย่าง n = 30 ได้ค่า R^2 ประมาณ 85% ขนาดตัวอย่าง n = 40 ได้ค่า R^2 ประมาณ 88% และ ขนาดตัวอย่าง n = 50 ได้ค่า R^2 ประมาณ 89%-90%

ตัวแบบถดถอยที่ศึกษาจะให้ค่า R^2 สูงที่สุดรองลงมา เมื่อทำการสุ่มตัวอย่าง 2 รอบ หรือ $m = 2$ ซึ่งที่ขนาดตัวอย่าง n = 10 ได้ค่า R^2 ประมาณ 46%-47% ขนาดตัวอย่าง n = 30 ได้ค่า R^2 ประมาณ 74%-77% ขนาดตัวอย่าง n = 40 ได้ค่า R^2 ประมาณ 77%-78% และ ขนาดตัวอย่าง n = 50 ได้ค่า R^2 ประมาณ 81%-82% และสำหรับการสุ่มตัวอย่างที่มากกว่า 2 รอบ ค่า R^2 มีแนวโน้มลดลงมาเรื่อย ๆ นั่นคือ เมื่อทำการสุ่มตัวอย่างที่จำนวนรอบ (m) เพิ่มมากขึ้น หรือขนาดชุดตัวอย่าง (k) น้อยลง จะทำให้ตัวแบบถดถอยที่ได้มีค่า R^2 ลดลง

เมื่อพิจารณาค่าความคลาดเคลื่อน (MSE) พบว่าที่ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 ตัวแบบถดถอยที่ศึกษาจะให้ค่าความคลาดเคลื่อนน้อยที่สุดเมื่อทำการสุ่มตัวอย่าง 1 รอบ หรือ $m = 1$ และค่า MSE จะลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น โดยส่วนใหญ่จะให้ค่า MSE น้อยที่สุดในกรณีที่ข้อมูลตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันที่ $\rho = 1.00$

จากผลการศึกษาดังกล่าวสามารถสรุปได้ว่า หากต้องการสร้างตัวแบบถดถอยภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล ควรทำการกำหนดขนาดตัวอย่างในการสุ่มตัวอย่างมาอย่างน้อย n=30 เนื่องจากว่าในการสุ่มตัวอย่างมาขนาด n=30 นั้นตัวแบบถดถอยให้ค่า R^2 ที่สูงมากพอ นั่นคือประมาณ 85% และเมื่อพิจารณาค่า MSE ร่วมด้วยพบว่า ตัวแบบถดถอยให้ค่า MSE ประมาณ 0.13-0.19 สำหรับในกรณีที่รอบของการสุ่มตัวอย่าง 1 รอบ หรือ $m=1$ และสำหรับในกรณีที่ทำการสุ่มตัวอย่างมี 2 รอบ หรือ $m=2$ ตัวแบบถดถอยให้ค่า R^2 ที่สูงมากพอเช่นกัน คือประมาณ 74%-77% ทั้งนี้ในการสร้างตัวแบบถดถอยควรคำนึงถึงข้อสมมติเบื้องต้นของการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายด้วย

5.3 อภิปรายผลการวิจัย

การวิจัยนี้ได้ศึกษาตัวแบบถดถอยเชิงเส้นอย่างง่ายที่สร้างขึ้นมาจากข้อมูลตัวอย่างที่ถูกสุ่มมาโดยการสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งตัวแบบที่ศึกษามีพารามิเตอร์คือ β_{0BRSS} และ β_{1BRSS} ซึ่งตัวประมาณค่าพารามิเตอร์ทั้งสองนี้เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียง และทำการจำลองข้อมูลตามแผนการจำลองข้อมูล เมื่อพิจารณาจากค่า R^2 ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 สำหรับกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันที่ $\rho=1.00$, $\rho=0.50$ และ $\rho=0.03$ ตัวแบบถดถอยที่ได้ให้ค่า R^2 สูงขึ้นเมื่อขนาดตัวอย่างเพิ่มมากขึ้น ซึ่งสอดคล้องกับการศึกษาของ Yaprak และ Alptekin ในปี 2007 ได้ศึกษาการประมาณพารามิเตอร์สำหรับตัวแบบถดถอยพหุคูณภายใต้การสุ่มตัวอย่างแบบชุดลำดับ นั่นคือเมื่อขนาดตัวอย่างเพิ่มมากขึ้น ค่า R^2 จะเพิ่มมากด้วย [Yaprak and Alptekin, 2007] ซึ่งผลที่ได้จากการศึกษานี้อาจเกิดจากการจำลองข้อมูลประชากรสำหรับตัวแปรอิสระและตัวแปรตาม ซึ่งได้ทำการจำลองข้อมูลประชากรภายใต้การแจกแจงปกติ ดังนั้นเมื่อทำการสุ่มตัวอย่างเพิ่มมากขึ้นจึงทำให้ข้อมูลตัวอย่างของทั้งตัวแปรอิสระและตัวแปรตามลู่เข้าสู่การแจกแจงแบบปกติด้วย ตามทฤษฎีแนวโน้มเข้าสู่ส่วนกลาง (CLT) จึงทำให้ตัวแปรอิสระและตัวแปรตามมีแนวโน้มของความสัมพันธ์กันเชิงเส้นตรงเช่นเดียวกันกับข้อมูลประชากรที่จำลองขึ้นมา และอีกทั้งในกระบวนการสุ่มตัวอย่างนั้นเป็นการสุ่มตัวอย่างขึ้นมาเป็นชุดตัวอย่าง แล้วมีการเรียงลำดับหน่วยตัวอย่างจากน้อยไปหามาก จากนั้นจะเลือกหน่วยตัวอย่างที่ถูกเรียงลำดับตามขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับมาเป็นตัวอย่าง จึงทำให้ข้อมูลของตัวแปรอิสระและตัวแปรตามมีแนวโน้มของความสัมพันธ์กันเชิงเส้นตรง นั่นคือตัวแบบถดถอยที่ได้จึงมีค่า R^2 เพิ่มขึ้นตามขนาดตัวอย่างที่สุ่มมา

เมื่อพิจารณาค่า R^2 ตามจำนวนรอบของการสุ่ม (m) ซึ่งขนาดชุดตัวอย่าง (k) จะขึ้นอยู่กับจำนวนรอบของการสุ่มตามแผนการจำลองข้อมูล พบว่าที่การสุ่มตัวอย่าง 1 รอบ หรือ $m=1$ สำหรับทุกระดับความสัมพันธ์ของข้อมูลประชากร ตัวแบบถดถอยที่ได้ให้ค่า R^2 สูงที่สุด ซึ่งค่า R^2 ที่สูงนั้นเมื่อทำการสุ่มตัวอย่าง $m=1$ เป็นการสุ่มตัวอย่างที่มีขนาดชุดตัวอย่างใหญ่ที่สุดในแต่ละขนาดตัวอย่าง เช่น ต้องการสุ่มตัวอย่างขนาด 30 หน่วย นั่นคือ ถ้าหาก $m=1$ ขนาดชุดตัวอย่าง $k=30$ ด้วยเหตุนี้ จึงทำให้ค่า R^2 สูงในกรณีที่ $m=1$ นอกจากนั้นอาจเกิดจากขั้นตอนในสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งเมื่อทำการสุ่มชุดตัวอย่าง $m=1$ ซึ่งเป็นชุดตัวอย่างขนาดใหญ่ที่สุดสำหรับขนาดตัวอย่าง n หน่วยมาแล้วนั้น ยังต้องนำหน่วยตัวอย่างมาเรียงลำดับจากน้อยไปหามากแล้วทำการเลือกหน่วยตัวอย่าง ด้วยเหตุนี้จึงทำให้ข้อมูลของทั้งตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันมากขึ้น จึงส่งผลทำให้ค่า R^2 สูงด้วย

เมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสอง (MSE) พบว่าที่ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 ตัวแบบถดถอยที่ศึกษาจะให้ค่าความคลาดเคลื่อนน้อยที่สุดเมื่อทำการสุ่มตัวอย่าง 1 รอบ หรือ $m=1$ และค่า MSE จะลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้นซึ่งผลการศึกษาที่ได้สอดคล้องกับ Al-Nasser และ Radaideh ในปี 2008 ซึ่งศึกษาการประมาณค่าตัวแบบถดถอยโดยใช้การสุ่มตัวอย่างแบบชุดลำดับ ซึ่งผลการศึกษาที่ได้พบว่า

ค่า MSE มีค่าน้อยที่สุดเมื่อขนาดตัวอย่างเพิ่มมากขึ้น หรือการสุ่มตัวอย่าง 1 รอบ และเมื่อประชากรมีค่า ρ มากพอ [Al-Nasser and Radaideh, 2008] ผลที่ได้ดังกล่าวอาจเนื่องมาจากขนาดตัวอย่างที่เพิ่มมากขึ้นทำให้ความคลาดเคลื่อนน้อยลง

5.4 ข้อเสนอแนะ

การวิจัยนี้เป็นการศึกษาการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อการสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย โดยได้สร้างตัวแบบขึ้นมา และทำการจำลองข้อมูลเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้สำหรับตัวแบบถดถอยอย่างง่ายที่สร้างขึ้นมา ซึ่งผลการศึกษาที่ได้เป็นเพียงการศึกษาโดยการจำลองข้อมูลเพียงอย่างเดียวเท่านั้น ดังนี้จึงควรมีการนำตัวแบบถดถอยที่ได้สร้างขึ้นมาไปใช้ในสถานการณ์จริงหรือใช้กับข้อมูลจริง อาทิเช่น การนำไปใช้ในทางการเกษตร หรือสร้างตัวแบบเพื่อพยากรณ์ เป็นต้น

เอกสารอ้างอิง

ชฎารัตน์ ถาปัน (2555). *การประมาณค่าแบบช่วงสำหรับค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์)*. มหาวิทยาลัยเชียงใหม่.

พิษณุ เจียวคุณ (2548). *การวิเคราะห์การถดถอย (Regression Analysis)*. ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่.

ยงยุทธ ไชยพงศ์ (2549). *เอกสารประกอบคำสอนวิชาการสำรวจตัวอย่าง*. ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่.

Alexander Levis. (2017). *Simulation and causal Inference methods for repeated measures or longitudinal data*. Department of Epidemiology, Biostatistics & Occupational Health McGill University Montreal. Canada.

Amjad D. Al-Nasser & Ahmed Radaideh (2008). *Estimation of simple linear regression model using L ranked set sampling*. Int. J. Open Problems Compt. Math., 1.

Bruce Ratner. (2009). *The correlation coefficient: Its values range between + 1 / − 1, or do they*. Journal of Targeting, Measurement and Analysis for Marketing; 17: 139-142.

Carmen R. Wilson Van Voorhis & Betsy L. Morgan. (2007). *Understanding power and ruled of Thumb for determining sample sizes*. Tutorials in Quantitative Methods for Psychology; vol.3(2): 43-50.

Dell T. R. & Clutter J. L.(1972). *Ranked set sampling theory with order statistics background*. Biometrics; 28:545-555.

George Casella and Roger L. Berger (2002). *Statistical Inference, 2nd edition*. United States of America. 542-543.

Hani M. Samawi & Faisal M. Ababneh. (2001). *On Regression analysis using ranked set sample*. Journal of statistical research. Volume 35, number 2, 93-105.

John A. Sokolowsak & Catherine M. Banks. (2009). *Principle of modeling and simulation*. Jonh Wiley & sons. Canada

Kowalczyk B. (2004). *Ranked set sampling and its applications in finite population studies.* Statistics in Transition; 6:1031-1046.

L.S. Halls & T. R. Dell (1966). *Trial of ranked set sampling for forage yields*. Forest Science; 12:22–26.

McIntyre G.A. (1952). *A method for unbiased selective sampling, using ranked set*. Austral. J. Agri. Rcs. 3: 385-390.

William G. Cochran (1997). *Sampling Techniques third edition*. The United States of America.

Yaprak Arzu & A. Alptekin Esin. (2007). *Parameter estimation in multiple linear regression models using ranked set sampling.* Commun.Fac.Sci.Univ.Series A1. Volume 56, number 1, 7-20.

Zehua Chen (2001). *Ranked-set sampling with regression-type estimation*. Elsevier, 181-192.

Zehua C. Zhidong Bai. Bimal K. Sinha (2003). *Ranked set sampling (Theory and Application)*. Springer.

ประวัติผู้วิจัย

1. ชื่อ - นามสกุล (ภาษาไทย) นางสาวชฎารัตน์ ถาปัน
ชื่อ - นามสกุล (ภาษาอังกฤษ) Miss Chadarat Tapan
สังกัดหน่วยงาน คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม
ประวัติการศึกษา ระดับปริญญาตรี วท.บ. สถิติ มหาวิทยาลัยแม่โจ้
ระดับปริญญาโท วท.ม. สถิติประยุกต์ มหาวิทยาลัยเชียงใหม่

2. ชื่อ - นามสกุล (ภาษาไทย) นางสาวสลิลทิพย์ แดงกองโค
ชื่อ - นามสกุล (ภาษาอังกฤษ) Miss Salinthip Daengkongkho
สังกัดหน่วยงาน คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม
ประวัติการศึกษา ระดับปริญญาตรี วท.บ. คณิตศาสตร์ มหาวิทยาลัยแม่โจ้
ระดับปริญญาโท วท.ม. คณิตศาสตร์ประยุกต์ มหาวิทยาลัยเชียงใหม่

ภาคผนวก

ภาพกิจกรรม

PROCEEDING

พะเยาวิจัย 8

PHAYAO RESEARCH CONFERENCE

24-25 มกราคม 2562

ณ มหาวิทยาลัยพะเยา จังหวัดพะเยา

ISBN: 978-616-7820-81-1

สารบัญ (ต่อ)

กลุ่มวิทยาศาสตร์และเทคโนโลยี การนำเสนอแบบ Poster Presentation (ต่อ)

การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

Development of Simple Linear Regression Model by Using Balance Ranked Set Sampling

สลิลทิพย์ แดงกองโค1 และชฎารัตน์ ถาปัน[1*]
Salinthip Daenkongkho1 and Chadarat Tapan[1*]

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling: BRSS) และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งได้ดำเนินการจำลองข้อมูลประชากรตามแผนการจำลองข้อมูล จากนั้นทำการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อสร้างตัวแบบถดถอย ซึ่งเกณฑ์ที่ใช้ในการพิจารณาตัวแบบถดถอยคือ ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: r^2) และค่าความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) ผลการศึกษาโดยสรุปพบว่า ตัวแบบถดถอยที่สร้างจากตัวอย่างที่มีขนาดตัวอย่างเพิ่มมากขึ้น และที่การสุ่มตัวอย่าง 1 รอบ (m=1) ได้ค่าสัมประสิทธิ์การตัดสินใจสูงที่สุด และได้ค่าความคลาดเคลื่อนกำลังสองน้อยที่สุด

คำสำคัญ: การสุ่มตัวอย่าง ตัวแบบถดถอยเชิงเส้น การสุ่มตัวอย่างแบบชุดลำดับ

Abstract

The purposes of this study are to propose simple linear regression model based on balance ranked set sampling and to check the accuracy and reliability of simple linear regression model. The investigation was done mainly via simulation study. Finite population was generated by simulation plans. Then, ranked set sampling was applied so as to obtain sample data to create a model. The criteria uses to consider the simple linear regression model are coefficient of determination (r^2) and mean square error (MSE). It was found that the simple linear regression model was generated from more sample size and cycle of sampling at 1 cycle (m=1) to obtain a most r^2 and at least MSE.

Keywords: sampling, linear regression model, ranked set sampling

[1] สาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม จังหวัดพิษณุโลก 65000
[1]Department of Mathematics, Faculty of Science and Technology, Pibulsongkram Rajabhat University, Phitsanulok Province 65000
*Corresponding author e-mail: n_nstat@hotmail.com

บทนำ

ในการจัดทำข้อสรุปสำหรับประชากรอันตะหรือประชากรจำกัด (Finite Population) เมื่อย้อนกลับไปในช่วงแรกเริ่มนั้นระเบียบวิธีที่ใช้ในการจัดทำข้อสรุปของประชากร คือการสำมะโน (Census) อันได้แก่การเก็บรวบรวมข้อมูลทุกหน่วยประชากร แล้วนำข้อมูลนั้นมาคำนวณค่าลักษณะประชากร (Population Characteristics) เช่น ค่าเฉลี่ยประชากร (Population Average) ค่ายอดรวมประชากร (Population Total) ค่าสัดส่วนประชากร (Population Proportion) หรือค่าอัตราส่วนประชากร (Population Ratio) [ชฎารัตน์ ถาปัน, 2555]

ด้วยข้อจำกัดของการสำมะโนไม่ว่าจะเป็นในส่วนของความทันสมัยของข้อมูล เวลาในการดำเนินการ บุคคลากรและงบประมาณในการเก็บข้อมูล ต่อมาได้มีการพัฒนาระเบียบวิธีการเก็บรวบรวมข้อมูลที่เรียกว่า การสำรวจตัวอย่าง ซึ่งจะทำการเก็บรวบรวมข้อมูลมาเพียงบางส่วนของประชากร ซึ่งในยุคแรกของการสำรวจตัวอย่างนั้นจะทำการสำรวจโดยใช้แผนการสุ่มตัวอย่างในลักษณะที่ไม่อิงความน่าจะเป็น (Non Probabilistic Sampling) แต่เนื่องด้วยแผนการสุ่มตัวอย่างที่ไม่อิงความน่าจะเป็นมีข้อจำกัดในส่วนของความน่าเชื่อถือได้ ต่อมาจึงได้มีการพัฒนาแผนการสุ่มตัวอย่างในแนวทางที่อิงความน่าจะเป็น (Probabilistic Sampling) จึงได้มีการพัฒนาเป็นทฤษฎีการสำรวจตัวอย่าง (Theory of Sample Survey) แผนการสุ่มตัวอย่างที่อิงความน่าจะเป็นประเภทแรก คือ แผนการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling)

หากแต่แผนการสุ่มตัวอย่างแบบง่ายเกิดข้อจำกัดทั้งในส่วนของความผิดพลาดในการประมาณค่าลักษณะประชากรที่สูงเมื่อประชากรที่ศึกษามีขนาดใหญ่ รวมทั้งข้อจำกัดในการบริหารจัดการเก็บข้อมูล (Data Collection) หรือเรื่องเวลาและค่าใช้จ่าย ต่อมาจึงได้มีแผนการสุ่มตัวอย่างที่ถูกพัฒนาขึ้นมาเพื่อแก้ปัญหาที่เกิดขึ้น อาทิเช่น แผนการสุ่มตัวอย่างแบบมีชั้นภูมิ (Stratified Sampling) และแผนการสุ่มตัวอย่างแบบกลุ่ม (Cluster Sampling) เป็นต้น แผนการสุ่มตัวอย่างดังกล่าวนั้นสามารถประมาณค่าลักษณะประชากรได้ หากแต่มีข้อจำกัดแตกต่างกันออกไป ซึ่งอาจส่งผลต่อความน่าเชื่อถือได้

ในปี 1952 McIntyre ได้เสนอวิธีการสุ่มตัวอย่างแบบชุดลำดับ (Ranked Set Sampling: RSS) เพื่อแก้ปัญหาการได้มาของแต่ละหน่วยตัวอย่างที่มีต้องใช้เวลานาน และมีค่าใช้จ่ายสูง ซึ่ง McIntyre ได้ใช้แผนการสุ่มตัวอย่างนี้ในการประมาณค่าเฉลี่ยของปริมาณหญ้าในพื้นที่หนึ่ง ซึ่งต้องใช้เวลานานในการเก็บเกี่ยวผลผลิตหญ้าทั้งหมด จากนั้นต้องใช้เวลาในการตากแห้งจึงทำการชั่งน้ำหนักเพื่อประมาณน้ำหนักหญ้าแห้ง ปัญหาดังกล่าวนี้ McIntyre ได้นำการสุ่มตัวอย่างแบบชุดลำดับมาแก้ปัญหาการเก็บข้อมูล [McIntyre G.A, 1952]

แผนการสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling) มีวิธีการดำเนินการโดยสุ่มตัวอย่างมา k ชุดตัวอย่าง ชุดตัวอย่างละ k หน่วย แล้วทำการเรียงลำดับหน่วยตัวอย่างในแต่ละชุดตัวอย่าง จากนั้นจึงทำการเลือกตัวอย่างจากทั้ง k ชุดตัวอย่าง มาเพียงชุดตัวอย่างละ 1 หน่วย โดยชุดตัวอย่างที่ 1 ทำการเลือกหน่วยตัวอย่างในลำดับที่ 1 ชุดตัวอย่างที่ 2 ทำการเลือกหน่วยตัวอย่างในลำดับที่ 2 จนกระทั่งชุดตัวอย่างที่ k เลือกหน่วยตัวอย่างลำดับที่ k เมื่อเลือกหน่วยตัวอย่างครบ k ชุดตัวอย่างแล้ว จะเรียกกระบวนการนี้ว่าการสุ่มตัวอย่าง 1 รอบ ซึ่งจะทำการสุ่มตัวอย่างเช่นนี้ไปจนครบ m รอบ แผนการสุ่มตัวอย่างนี้ทำให้สามารถลดขนาดตัวอย่างในการเก็บรวบรวม และได้ค่าประมาณลักษณะประชากรที่น่าเชื่อถือได้เมื่อเปรียบเทียบกับแผนการสุ่มตัวอย่างแบบง่ายดังการศึกษาของ Dell และ Clutter [Dell and Clutter, 1972]

ในทางปฏิบัติสำหรับการนำข้อมูลที่ได้จากการสำรวจตัวอย่างไปหาค่าประมาณของค่าลักษณะประชากรแล้วนั้น ยังมีการนำข้อมูลที่เก็บรวบรวมไปทำการวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูล รวมไปถึงการสร้างตัวแบบเพื่อพยากรณ์ ซึ่งตัวแบบที่นิยมใช้กันอย่างแพร่หลายคือ ตัวแบบถดถอยเชิงเส้น (Linear Regression Model) ซึ่งแบ่ง

ออกเป็นตัวแบบถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Model) และตัวแบบถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Model)

ตัวแบบถดถอยเชิงเส้นอย่างง่ายเป็นตัวแบบที่วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระ 1 ตัวแปร กับตัวแปรตาม 1 ตัวแปร โดยตัวแปรทั้งสองมีความสัมพันธ์กันเชิงเส้นตรง ซึ่งมีตัวแบบถดถอยเชิงเส้นอย่างง่ายของประชากร คือ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ซึ่งมี X_i เป็นตัวแปรอิสระ Y_i เป็นตัวแปรตาม β_0, β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า และ ε_i เป็นความคลาดเคลื่อน

ในการสร้างตัวแบบถดถอยเชิงเส้นนั้นหากข้อมูลที่ถูกนำมาวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างตัวแปรนั้นเป็นข้อมูลที่สามารถอธิบายคุณลักษณะประชากรได้เป็นอย่างดีหรืออธิบายได้ใกล้เคียงกับประชากรแล้วนั้น ตัวแบบพยากรณ์หรือตัวแบบของความสัมพันธ์ระหว่างตัวแปรนั้นย่อสามารถพยากรณ์ค่าที่ต้องการได้อย่างถูกต้องใกล้เคียงกับความเป็นจริง ดังนั้นในการวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

วิธีการศึกษา

การวิจัยนี้ได้ศึกษาในส่วนของการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ตัวแบบถดถอยเชิงเส้นอย่างง่าย และตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล รวมทั้งการจำลองข้อมูลเพื่อตรวจสอบความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งมีเนื้อหาดังต่อไปนี้

1. การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling: BRSS)

การสำรวจตัวอย่างเพื่อเก็บรวบรวมข้อมูลในกรณีที่ประชากรที่ศึกษามีขนาดใหญ่จึงทำให้ต้องใช้เวลารวมถึงค่าใช้จ่ายสำหรับการศึกษาเป็นจำนวนมาก ในปี 1952 McIntyre จึงได้พัฒนาวิธีการสุ่มตัวอย่างขึ้นมาเพื่อหาวิธีการสุ่มตัวอย่างที่สามารถลดค่าใช้จ่าย (Cost-effective Sampling Method) และประหยัดเวลาในการดำเนินการ ซึ่งในเวลาต่อมาเรียกวิธีการสุ่มตัวอย่างดังกล่าวว่า การสุ่มตัวอย่างแบบชุดลำดับ [McIntyre G.A., 1952]

การสุ่มตัวอย่างแบบชุดลำดับมีขั้นตอนการสุ่มตัวอย่างที่ซับซ้อนมากกว่าการสุ่มตัวอย่างแบบง่าย นั่นคือการสุ่มตัวอย่างแบบง่ายจะทำการสุ่มตัวอย่างขึ้นมาเพียง n หน่วยเท่านั้นจากประชากรขนาด N หน่วย แต่สำหรับการสุ่มตัวอย่างแบบชุดลำดับจะทำการสุ่มตัวอย่างมา n_r หน่วย และ $r = 1,2,\ldots,k$ เรียกว่า ชุดตัวอย่าง ซึ่งจะทำการสุ่มตัวอย่างขึ้นมา k ชุดตัวอย่าง เรียกว่าการสุ่มตัวอย่าง 1 รอบ และกำหนดให้ $Y_{[r]i'}$ คือ หน่วยตัวอย่างลำดับที่ r ของชุดตัวอย่างที่ i' สำหรับการสุ่มตัวอย่าง 1 รอบ ได้แสดงดังนี้

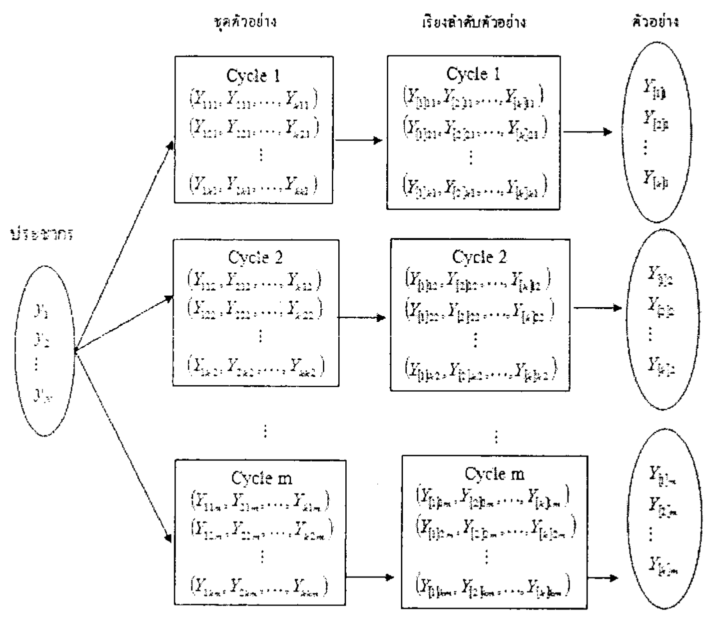
$$\begin{array}{cccc} Y_{[1]1} & Y_{[1]2} & \cdots & Y_{[1]n_1} \\ Y_{[2]1} & Y_{[2]2} & \cdots & Y_{[2]n_2} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{[k]1} & Y_{[k]2} & \cdots & Y_{[k]n_k} \end{array}$$

ถ้าขนาดตัวอย่างในการสุ่มแต่ละชุดตัวอย่างเท่ากัน คือ $n_1 = n_2 = \ldots = n_k$ แล้วจะเรียกการสุ่มตัวอย่างแบบชุดลำดับนี้ว่า การสุ่มตัวอย่างแบบชุดลำดับได้ดุล (Balance Ranked Set Sampling) [Zehua C. Zhidong Bai. & Bimal K. Sinha, 2003]

ขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเริ่มต้นด้วยการสุ่มตัวอย่างแบบง่ายไม่ใส่คืนขนาด k หน่วย จากประชากรขนาด N หน่วย ซึ่งจะเรียกว่าชุดตัวอย่างที่ 1 แล้วนำหน่วยตัวอย่าง (Sampling Units) ทั้ง k หน่วยมาเรียงลำดับจากน้อยไปหามากด้วยการคาดคะเนหรือการประมาณด้วยสายตา (Judgment Ranking) จากนั้นหน่วยตัวอย่างที่อยู่ในลำดับที่ 1 จะถูกเลือกเป็นตัวอย่าง และทำการวัดค่าของตัวแปรที่สนใจ ส่วนหน่วยตัวอย่างที่เหลือถูกตัดทิ้งหรือไม่สนใจ

เมื่อสิ้นสุดการสุ่มตัวอย่างชุดที่ 1 แล้ว ต่อจากนั้นสุ่มตัวอย่างแบบง่ายไม่ใส่คืนขนาด k หน่วย จากประชากรขนาด N หน่วย ซึ่งนั่นหมายความว่า หน่วยประชากรที่ถูกสุ่มเป็นตัวอย่างในชุดตัวอย่างที่ 1 ถูกนำกลับมาใส่ในประชากรอีกครั้ง หน่วยประชากร k หน่วยที่ถูกสุ่มในครั้งที่ 2 เรียกว่าชุดตัวอย่างที่ 2 จากนั้นนำหน่วยตัวอย่างมาเรียงลำดับจากน้อยไปหามาก หน่วยตัวอย่างที่อยู่ในลำดับที่ 2 ในชุดตัวอย่างที่ 2 นี้จะถูกเลือกเป็นตัวอย่าง แล้วทำการวัดค่าของตัวแปรที่สนใจ และหน่วยที่เหลือถูกตัดทิ้งหรือไม่สนใจ

กระบวนการสุ่มตัวอย่างนี้ดำเนินการต่อไปเรื่อย ๆ จนกระทั่งสุ่มตัวอย่างไปจนครบ k ชุดตัวอย่าง และจะเรียกกระบวนการสุ่มตัวอย่างจนครบ k ชุดตัวอย่างนี้ว่า "รอบ (Cycle)" ซึ่งจะทำการสุ่มตัวอย่างทั้งหมด m รอบ และการสุ่มหน่วยตัวอย่างในแต่ละชุดตัวอย่างที่ทำการสุ่มขึ้นมาทั้ง k ชุดตัวอย่างในแต่ละรอบนั้น การได้หน่วยตัวอย่าง k หน่วย ในแต่ละชุดตัวอย่างไม่มีผลกระทบต่อการได้หน่วยตัวอย่างในชุดตัวอย่างอื่น ๆ [Kowalczyk B., 2004] ขั้นตอนการสุ่มตัวอย่างแสดงในรูปที่ 1



ภาพที่ 1 แสดงขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

จากภาพที่ 1 เป็นการสุ่มตัวอย่างแบบชุดลำดับได้ดุลมา m รอบ รอบละ k ชุดตัวอย่าง หน่วยตัวอย่างที่ได้คือ $Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[k]1}, \ldots, Y_{[1]m}, Y_{[2]m}, \ldots, Y_{[k]m}$ มีขนาด mk หน่วย

2. ตัวแบบถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Model)

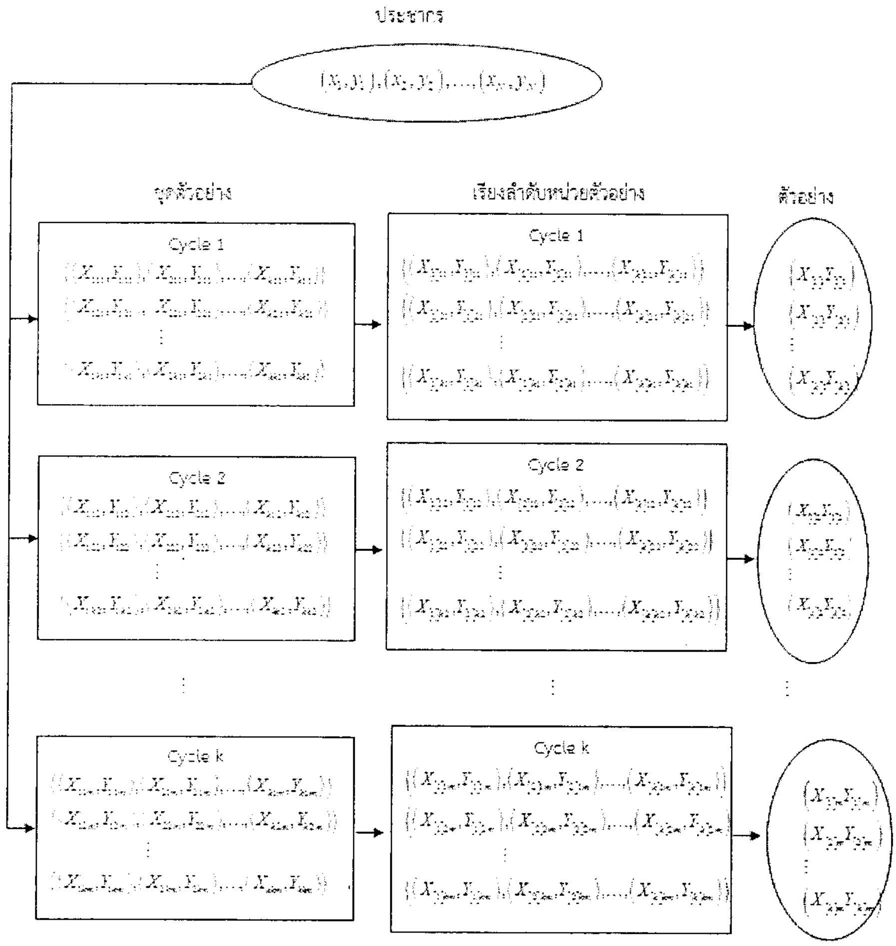
ตัวแบบถดถอยเชิงเส้นอย่างง่ายเป็นการศึกษาความสัมพันธ์ระว่างตัวแปร 2 ตัว ซึ่งประกอบไปด้วยตัวแปรอิสระ (Independence Variable) 1 ตัว และตัวแปรตาม (Dependence Variable) 1 ตัว ซึ่งความสัมพันธ์ระหว่าง 2 ตัวแปรนี้เป็นไปในลักษณะความสัมพันธ์เชิงเส้นตรง โดยมีตัวแบบการถดถอยสำหรับประชากร (Population Regression Model) คือ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ โดยที่ Y_i เป็นค่าสังเกตของตัวแปรตาม β_0 เป็นพารามิเตอร์แทนระยะตัดแกน Y

β_1 เป็นพารามิเตอร์แทนความชันของเส้นถดถอย X_i เป็นค่าสังเกตของตัวแปรอิสระ(ค่าคงที่ที่ทราบค่า) และ ε_i เป็นค่าความคลาดเคลื่อนเชิงสุ่ม

สำหรับการประมาณค่าพารามิเตอร์ β_0 และ β_1 จะใช้วิธีกำลังสองน้อยสุด (Least Square Method) ตัวประมาณที่ได้คือ $\beta_0 = \bar{Y} - \beta_1\bar{X}$ และ $\hat{\beta}_1 = \dfrac{n\left(\sum_{i=1}^{n} X_iY_i\right) - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$ ซึ่งตัวประมาณดังกล่าวมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียง และความแปรปรวนต่ำสุด

3. การวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างถูกสุ่มตัวอย่างแบบชุดลำดับดับได้ดุล

ในปี 2001 Hani M. Samuwi และ Faisal M. Ababneh ได้ศึกษาการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายโดยใช้ตัวอย่างจากการสุ่มตัวอย่างแบบชุดลำดับ [Hani M. Samuwi and Faisal M. Ababneh, 2001] และในปี 2002 Chen ได้แนะนำตัวประมาณค่าเฉลี่ยประชากรของตัวแบบถดถอยสำหรับการสุ่มตัวอย่างแบบชุดลำดับ ซึ่งได้ทำการเรียงลำดับหน่วยตัวอย่าง (ตัวแปรตาม) โดยตัวแปรอิสระ จากนั้นทำการวัดค่าตัวแปรอิสระและตัวแปรตามสำหรับการประมาณค่าเฉลี่ยประชากร [Chen Z., 2001] สำหรับการวิจัยนี้ในการสุ่มตัวอย่างจากประชากรที่มีความสัมพันธ์กันหรือมีข้อมูลประชากรสำหรับตัวแปรตาม (Y) และตัวแปรอิสระ (X) ในที่นี้ได้ศึกษาการสุ่มตัวอย่างจากข้อมูลประชากรทั้งตัวแปรตามและตัวแปรอิสระด้วยการสุ่มตัวอย่างแบบชุดลำดับได้ดุลเพื่อนำมาสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย ขั้นตอนการสุ่มตัวอย่างสำหรับสร้างตัวแบบถดถอยดังแสดงในภาพที่ 2



ภาพที่ 2 ขั้นตอนการสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับสร้างตัวแบบถดถอย

จากภาพที่ 2 แสดงให้เห็นถึงการสุ่มหน่วยประชากรแล้วนำมาเรียงลำดับสำหรับตัวแปรอิสระ ($X_{[k]m}$) และตัวแปรตาม($Y_{[k]m}$) มา m รอบ รอบละ k ชุดตัวอย่าง จากนั้นในแต่ละชุดตัวอย่างหน่วยตัวอย่างทั้ง k หน่วยถูกเรียงลำดับจากน้อยไปหามาก สำหรับชุดตัวอย่างที่ 1 ของรอบที่ 1 หน่วยตัวอย่างที่อยู่ในลำดับที่ 1 จะถูกเลือกเป็นตัวอย่าง จากนั้น ในชุดตัวอย่างที่ 2 ของรอบที่ 1 หน่วยตัวอย่างที่อยู่ในลำดับที่ 2 จะถูกเลือกเป็นตัวอย่าง จนกระทั่งในชุดตัวอย่างที่ k ของรอบที่ m หน่วยตัวอย่างที่อยู่ในลำดับที่ k จะถูกเลือกมาเป็นตัวอย่าง ดังนั้นจากสุ่มตัวอย่างแบบชุดลำดับได้ดุลสำหรับการวิเคราะห์ถดถอยเชิงเส้นอย่างง่ายซึ่งประชากรมีความสัมพันธ์กัน จะได้ตัวอย่างขนาด $n = mk$ หน่วย จากประชากรขนาด N หน่วย คือ

$$\begin{array}{c}
\left(X_{[1]1}Y_{[1]1}\right),\left(X_{[2]1}Y_{[2]1}\right),\ldots,\left(X_{[k]1}Y_{[k]1}\right) \\
\left(X_{[1]2}Y_{[1]2}\right),\left(X_{[2]2}Y_{[2]2}\right),\ldots,\left(X_{[k]2}Y_{[k]2}\right) \\
\vdots \qquad\qquad\qquad\qquad \vdots \\
\left(X_{[1]m}Y_{[1]m}\right),\left(X_{[2]m}Y_{[2]m}\right),\ldots,\left(X_{[k]m}Y_{[k]m}\right)
\end{array}$$

ตัวแบบพยากรณ์สำหรับประชากรที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งตัวแบบประชากรที่ศึกษา คือ $Y_{[r]iBRSS} = \beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} + \varepsilon_{[r]i}$; $r = 1,2,3,\ldots,k$ $i = 1,2,3,\ldots,m$ เมื่อ β_{0BRSS} และ β_{1BRSS} คือ พารามิเตอร์ $\varepsilon_{[r]i}$ คือความคลาดเคลื่อน $Y_{[r]iBRSS}$ คือ ตัวแปรตามที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล และ $X_{[r]iBRSS}$ คือ ตัวแปรอิสระที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล

และตัวแบบของตัวอย่างสำหรับพยากรณ์ค่าเฉลี่ยประชากร $\left(\hat{Y}_{[r]iBRSS}\right)$ คือ

$\hat{Y}_{[r]iBRSS} = \hat{\beta}_{0BRSS} + \hat{\beta}_{1BRSS}X_{[r]iBRSS}$ เมื่อ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ คือ ตัวประมาณพารามิเตอร์ $\hat{Y}_{[r]iBRSS}$ คือตัวประมาณตัวแปรตาม และ $X_{[r]iBRSS}$ คือตัวแปรอิสระ สำหรับการประมาณค่าพารามิเตอร์ β_{0BRSS} และ β_{1BRSS} จะใช้วิธีกำลังสองน้อยที่สุด (Least – Squares Estimates) ตัวประมาณที่ได้คือ

$$\hat{\beta}_{1BRSS} = \frac{mk\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}Y_{[r]iBRSS} - \sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}\sum_{i=1}^{m}\sum_{r=1}^{k} Y_{[r]iBRSS}}{mk\sum_{i=1}^{m}\sum_{r=1}^{k} X^2_{[r]iBRSS} - \left(\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}\right)^2}$$ และ

$$\hat{\beta}_{0BRSS} = \frac{\sum_{i=1}^{m}\sum_{r=1}^{k} Y_{[r]iBRSS}}{mk} - \hat{\beta}_{1BRSS}\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}}{mk}\right)$$

ตัวประมาณ $\hat{\beta}_{0BRSS}$ เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียง

$$E\left(\hat{\beta}_{0BRSS}\right) = E\left[\frac{\sum_{i=1}^{m}\sum_{r=1}^{k} Y_{[r]iBRSS}}{mk} - \hat{\beta}_{1BRSS}\left(\frac{\sum_{i=1}^{m}\sum_{r=1}^{k} X_{[r]iBRSS}}{mk}\right)\right]$$

$$= \beta_{0BRSS}$$

และตัวประมาณพารามิเตอร์ $\hat{\beta}_{1BRSS}$ เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียงเช่นกัน

ในการศึกษานี้ได้ทำการจำลองข้อมูลเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้ตัวอย่างที่ถูกสุ่มมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งมีแผนการจำลองข้อมูลดังนี้

1. ประชากรที่ศึกษาสร้างขึ้นมาภายใต้การแจกแจงปกติ (Normal Distribution) โดยกำหนดให้ข้อมูลประชากรมีความสัมพันธ์กัน นั่นคือข้อมูลประชากรตัวแปรอิสระ (x) และตัวแปรตาม (y) มีความสัมพันธ์กัน

2. กำหนดค่าสหสัมพันธ์ (Correlations) สำหรับตัวแปรอิสระและตัวแปรตาม แบ่งออกเป็น 3 ระดับความสัมพันธ์ ได้แก่ ระดับความสัมพันธ์มาก คือ 0.71-1.00 ระดับความสัมพันธ์ปานกลาง คือ 0.31-0.70 และระดับความสัมพันธ์ต่ำ คือ 0.00-0.30 [Bruce Ratner, 2009]

2. ขนาดประชากร (N) ที่ศึกษาเท่ากับ 1,000

3. ขนาดของตัวอย่าง (n) ที่ศึกษา คือ n = 5, 10, 30, 40, 50 โดยที่ $n = mk$ ซึ่ง k คือ ขนาดตัวอย่างในการสุ่มแต่ละชุดตัวอย่างและเป็นจำนวนชุดตัวอย่างของการสุ่มแต่ละรอบ และ m คือ จำนวนรอบในการสุ่มตัวอย่าง

4. ระดับความเชื่อมั่นที่ศึกษา คือ 95% และ 99%

5. ทำการทดลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ที่กำหนด

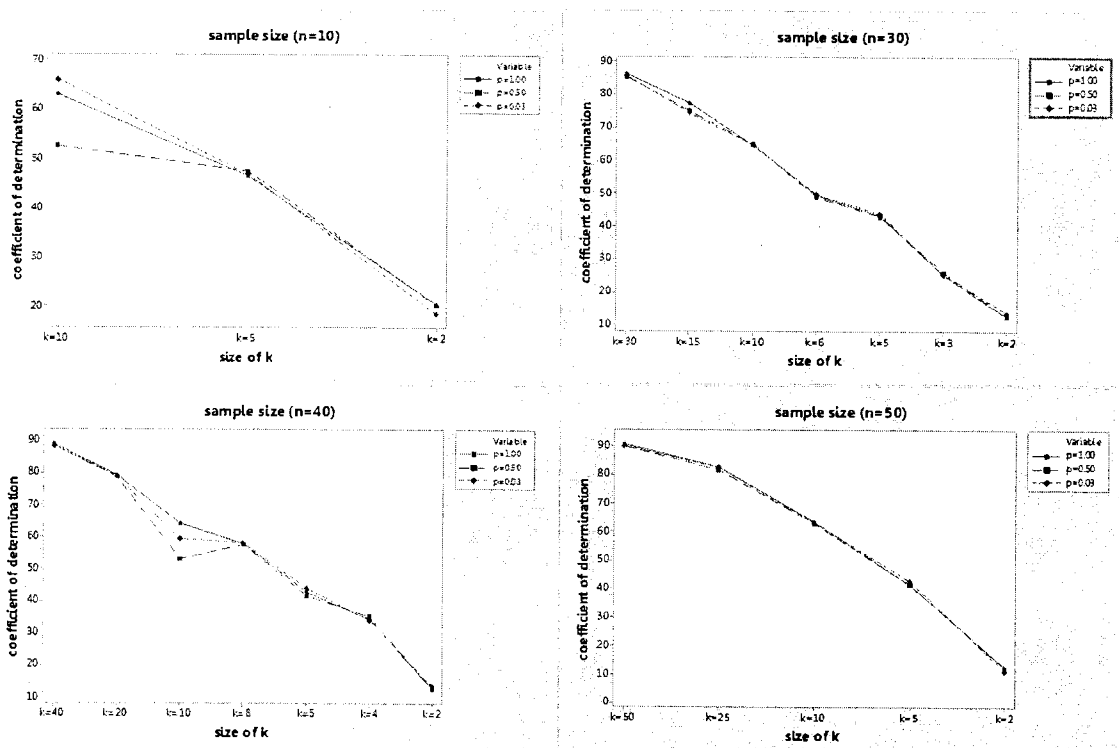
ผลการศึกษา

การวิจัยนี้ได้ศึกษาตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งตัวแบบถดถอยที่ศึกษาคือ $Y_{[r]iBRSS} = \beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} + \varepsilon_{[r]i}$ และตัวประมาณค่าสำหรับพารามิเตอร์คือ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ ดังได้กล่าวมาแล้วข้างต้น และเมื่อทำการตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล จากการจำลองข้อมูลตามแผนการจำลองข้อมูลได้แสดงผลดังตารางที่ 1 และตารางที่ 2

ตารางที่ 1 ค่าสัมประสิทธิ์การตัดสินใจ (r^2) สำหรับกรณีต่าง ๆ ตามแผนการจำลองข้อมูล

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	r^2 (%)		
			$\rho=1.00$	$\rho=0.50$	$\rho=0.03$
5	1	5	52.813	52.524	52.589
10	1	10	62.916	66.456	65.806
	2	5	46.477	47.433	46.818
	5	2	20.252	20.210	18.459
30	1	30	85.994	85.292	85.464
	2	15	77.163	74.969	74.286
	3	10	64.648	64.893	64.488
	5	6	49.445	48.696	49.815
	6	5	43.138	42.793	43.739
	10	3	25.104	25.910	25.259
	15	2	12.333	12.418	13.522
40	1	40	88.900	88.251	88.348
	2	20	79.376	78.702	79.122
	4	10	64.332	52.889	59.257
	5	8	57.726	57.483	58.236
	8	5	42.900	41.552	43.856
	10	4	34.425	35.194	34.120
	20	2	12.619	12.671	13.347
50	1	50	90.499	89.650	89.871
	2	25	82.707	81.446	82.453
	5	10	63.346	62.656	63.070
	10	5	41.178	41.100	42.550
	25	2	12.583	11.973	11.235

จากตารางที่ 1 พบว่าข้อมูลประชากรที่กำหนดค่าสหสัมพันธ์ (ρ) สำหรับตัวแปรอิสระและตัวแปรตามที่แตกต่างกันทั้ง 3 ระดับนั้น เมื่อนำตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย พบว่า ค่า r^2 ที่ได้จากตัวแบบถดถอยเชิงเส้นในแต่ละกรณีสำหรับตัวอย่างขนาด 5 10 30 40 และ 50 มีค่าใกล้เคียงกันเมื่อค่าสหสัมพันธ์ต่างกัน แต่ที่ $\rho = 1.00$ อาจมีค่า r^2 สูงกว่า $\rho = 0.50$ และ $\rho = 0.03$ เล็กน้อย และที่ n = 5 มีค่าสัมประสิทธิ์การตัดสินใจประมาณ 50% ส่วนที่ n =10, 30, 40 และ 50 ที่จำนวนการสุ่ม 1 รอบ ($m = 1$) มีค่า r^2 สูงที่สุด คือประมาณ 65%, 85%, 88% และ 90% ตามลำดับ เมื่อนำผลการศึกษาที่ได้ในตารางที่ 2 มาพลอต กราฟแสดงได้ดังรูปที่ 3 สำหรับกรณีที่ n = 10, 30, 40 และ 50



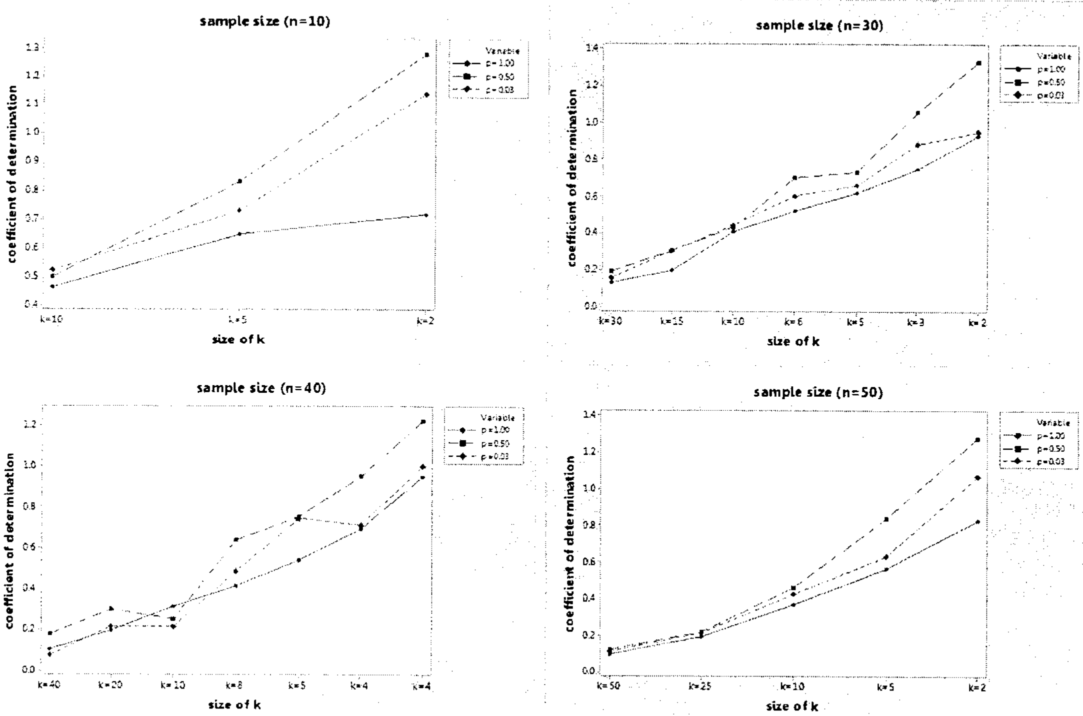
ภาพที่ 3 แสดงกราฟระหว่างค่าสัมประสิทธิ์การตัดสินใจ (r^2) กับขนาดชุดตัวอย่าง (k)

ตารางที่ 2 ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) สำหรับกรณีต่าง ๆ ตามแผนการจำลองข้อมูล

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	MSE		
			$\rho=1.00$	$\rho=0.50$	$\rho=0.03$
5	1	5	0.527	0.897	0.989
10	1	10	0.459	0.497	0.519
	2	5	0.649	0.833	0.732
	5	2	0.718	1.280	1.140
30	1	30	0.131	0.191	0.154
	2	15	0.198	0.308	0.307
	3	10	0.407	0.434	0.444

ขนาดตัวอย่าง (n)	จำนวนรอบ (m)	ขนาดชุดตัวอย่าง (k)	MSE		
			$\rho=1.00$	$\rho=0.50$	$\rho=0.03$
	5	6	0.523	0.702	0.607
	6	5	0.619	0.732	0.661
	10	3	0.751	1.060	0.885
	15	2	0.928	1.333	0.952
40	1	40	0.102	0.177	0.076
	2	20	0.195	0.298	0.216
	4	10	0.314	0.251	0.215
	5	8	0.416	0.638	0.485
	8	5	0.541	0.753	0.742
	10	4	0.692	0.951	0.712
	20	2	0.948	1.224	1.002
50	1	50	0.093	0.120	0.111
	2	25	0.192	0.216	0.211
	5	10	0.370	0.460	0.424
	10	5	0.565	0.834	0.630
	25	2	0.824	1.278	1.070

จากตารางที่ 2 พบว่าที่ข้อมูลประชากรที่กำหนดค่าสหสัมพันธ์ (ρ) สำหรับตัวแปรอิสระและตัวแปรตามที่แตกต่างกันทั้ง 3 ระดับนั้น เมื่อนำตัวอย่างที่ได้จากการสุ่มตัวอย่างแบบชุดลำดับได้ดุลไปสร้างตัวแบบถดถอยเชิงเส้นอย่างง่าย พบว่า ที่ n = 5 และ ρ =1.00 ตัวแบบที่ได้ให้ค่าความคลาดเคลื่อนกำลังสองน้อยที่สุด ส่วนที่ n = 10, 30, 40, 50 ที่ m = 1 ตัวแบบที่ได้ให้ค่าความคลาดเคลื่อนกำลังสองน้อยที่สุดสำหรับทุกค่าสหสัมพันธ์ (ρ) และจะมีค่าความคลาดเคลื่อนกำลังสองเพิ่มขึ้นเมื่อมีจำนวนรอบของการสุ่มตัวอย่าง (m)เพิ่มขึ้น หรือขนาดชุดตัวอย่าง (k) ลดลง เมื่อนำผลการศึกษาที่ได้ในตารางที่ 2 มาพลอตกราฟแสดงได้ดังรูปที่ 4 สำหรับกรณีที่ n = 10, 30, 40 และ 50



***ภาพที่* 4** แสดงกราฟระหว่างค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) กับขนาดชุดตัวอย่าง (k)

อภิปรายและสรุปผล

อภิปรายผลการศึกษา

การวิจัยนี้ได้ศึกษาตัวแบบถดถอยเชิงเส้นอย่างง่ายที่สร้างขึ้นมาจากข้อมูลตัวอย่างที่ถูกสุ่มมาโดยการสุ่มตัวอย่างแบบชุดลำดับได้ดุลซึ่งตัวแบบที่ศึกษามีพารามิเตอร์คือ β_{0BRSS} และ β_{1BRSS} ซึ่งตัวประมาณค่าพารามิเตอร์ทั้งสองนี้เป็นตัวประมาณที่มีคุณสมบัติไม่เอนเอียง และเมื่อทำการจำลองข้อมูลตามแผนการจำลองข้อมูลพบว่าผลการจำลองข้อมูลสอดคล้องกับการศึกษาของ Zehua Chen ซึ่งได้ศึกษาเกี่ยวกับการสุ่มตัวอย่างแบบชุดลำดับสำหรับการประมาณค่าพารามิเตอร์ นั่นคือที่ค่า ρ เพิ่มมากขึ้นทำให้ตัวแบบที่ได้มีค่าสหสัมพันธ์ (Relative Efficient: r) เพิ่มมากขึ้น [Zehua Chen, 2001] เช่นเดียวกับการวิจัยนี้ที่ค่า ρ อยู่ในระดับสูงค่า r^2 ของตัวแบบที่ได้โดยรวมมีค่ามากกว่าตัวแบบที่ตัวอย่างถูกสุ่มมาจากประชากรที่มีค่า ρ อยู่ในระดับน้อยกว่า จะเห็นได้ว่าผลที่ได้อาจจะเกิดจากข้อมูลตัวอย่างที่ถูกสุ่มมาจากประชากรที่มีความสัมพันธ์ที่ต่างกัน

และเมื่อพิจารณาค่า r^2 ตามขนาดชุดตัวอย่าง (k) พบว่าที่การสุ่มตัวอย่าง 1 รอบ ($m=1$) สำหรับทุกระดับความสัมพันธ์ของข้อมูลประชากร และทุกขนาดตัวอย่างตัวแบบที่ได้ให้ค่า r^2 สูงที่สุด ซึ่งผลที่ได้อาจเกิดจากการสุ่มตัวอย่างในแต่ละชุดตัวอย่างแล้วนำมาเรียงลำดับจึงทำให้ข้อมูลตัวอย่างสำหรับตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันมากกว่าการสุ่มตัวอย่างที่ชุดตัวอย่างขนาดเล็กกว่า หรือจำนวนการสุ่มตัวอย่างที่มากกว่า 1 รอบ

และเมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสองพบว่ามีผลการศึกษาที่สอดคล้องกับ Amjad D. Al-Nasser และ Ahmed Radaideh ซึ่งศึกษาการประมาณค่าตัวแบบถดถอยโดยใช้การสุ่มตัวอย่างแบบชุดลำดับ L ซึ่งผลการศึกษาที่ได้พบว่าค่า MSE มีค่าน้อยที่สุดเมื่อขนาดตัวอย่างเพิ่มมากขึ้น หรือการสุ่มตัวอย่าง 1 รอบ และเมื่อประชากรมีค่า ρ มากพอ [Amjad D. Al-Nasser and Ahmed Radaideh, 2008] ผลที่ได้ดังกล่าวอาจเนื่องมาจากขนาดตัวอย่างที่เพิ่มมากขึ้นทำให้ความคลาดเคลื่อนน้อยลง

สรุปผลการศึกษา

การวิจัยนี้มีวัตถุประสงค์คือ สร้างตัวแบบถดถอยเชิงเส้นอย่างง่ายภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล ซึ่งตัวแบบถดถอยคือ $Y_{[r]iBRSS} = \beta_{0BRSS} + \beta_{1BRSS}X_{[r]iBRSS} + \varepsilon_{[r]i}$ โดยมีตัวประมาณค่าพารามิเตอร์ คือ $\hat{\beta}_{0BRSS}$ และ $\hat{\beta}_{1BRSS}$ และเพื่อตรวจสอบความถูกต้องและความน่าเชื่อถือได้ของตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อตัวอย่างมาจากการสุ่มตัวอย่างแบบชุดลำดับได้ดุล พบว่าที่ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 ตัวแบบที่ศึกษาจะให้ค่า r^2 สูงสุดเมื่อทำการสุ่มตัวอย่าง 1 รอบ ($m=1$) และสำหรับค่าความคลาดเคลื่อน (MSE) พบว่าที่ขนาดตัวอย่าง n = 5, 10, 30, 40 และ 50 ตัวแบบที่ศึกษาจะให้ค่าความคลาดเคลื่อนน้อยที่สุดเมื่อทำการสุ่มตัวอย่าง 1 รอบ ($m=1$)

กิตติกรรมประกาศ

ขอขอบคุณ กองทุนพัฒนาการวิจัยและบริหารจัดการงานวิจัย มหาวิทยาลัยราชภัฏพิบูลสงคราม ทุนวิจัยเพื่อพัฒนานักวิจัย ประจำปีงบประมาณ พ.ศ. 2561 ซึ่งเป็นแหล่งให้ทุนสำหรับการศึกษาวิจัยในครั้งนี้

เอกสารอ้างอิง

ชฎารัตน์ ถาปัน (2555). **การประมาณค่าแบบช่วงสำหรับค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล.** (วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์). มหาวิทยาลัยเชียงใหม่.

Amjad D. Al-Nasser & Ahmed Radaideh (2008). **Estimation of simple linear regression model using L ranked set sampling.** Int. J. Open Problems Compt. Math., 1.

Bruce Ratner. (2009). The correlation coefficient: Its values range between + 1 / – 1, or do they. **Journal of Targeting, Measurement and Analysis for Marketing**; 17: 139–142.

Dell T. R. & Clutter J. L.(1972). **Ranked set sampling theory with order statistics background**. Biometrics; 28:545–555.

Hani M. Samawi & Faisal M. Ababneh. (2001). On Regression analysis using ranked set sample. **Journal of statistical research**. Volume 35, number 2, 93–105.

Kowalczyk B. (2004). **Ranked set sampling and its applications in finite population studies.** Statistics in Transition; 6:1031–1046.

L.S. Halls & T. R. Dell (1966). **Trial of ranked set sampling for forage yields.** Forest Science; 12:22–26.McIntyre G.A. (1952). **A method for unbiased selective sampling, using ranked set.** Austral. J. Agri. Rcs. 3: 385–390.

William G. Cochran (1997). **Sampling Techniques third edition.** The United States of America.Yaprak Arzu & A. Alptekin Esin. (2007). **Parameter estimation in multiple linear regression models using ranked set sampling.** Commun.Fac.Sci.Univ.Series A1. Volume 56, number 1, 7–20.

Zehua Chen (2001). **Ranked-set sampling with regression-type estimation.** Elsevier, 181–192.

Zehua C. Zhidong Bai. Bimal K. Sinha (2003). **Ranked set sampling (Theory and Application).** Springer.

พะเยาวิจัย 8
PHAYAO RESEARCH CONFERENCE 2019

มหาวิทยาลัยพะเยา

ขอมอบเกียรติบัตรเพื่อรับรองว่าผลงานวิจัย

เรื่อง การพัฒนาตัวแบบถดถอยเชิงเส้นอย่างง่ายเมื่อใช้การสุ่มตัวอย่างแบบชุดลำดับได้ดุล

โดย ชฎารัตน์ ถาปัน

ได้ผ่านการพิจารณาจากคณะกรรมการผู้ทรงคุณวุฒิ และได้นำเสนอผลงานประเภท Poster Presentation
กลุ่มการวิจัยวิทยาศาสตร์และเทคโนโลยี (ด้านวิทยาศาสตร์ 1) วันพฤหัสบดีที่ 24 มกราคม 2562
ในการประชุมทางวิชาการระดับชาติพะเยาวิจัย ครั้งที่ 8 ระหว่างวันที่ 24 - 25 มกราคม พ.ศ. 2562
ณ หอประชุมพญางำเมือง มหาวิทยาลัยพะเยา จังหวัดพะเยา
ให้ไว้ ณ วันที่ 25 มกราคม พ.ศ. 2562

(รองศาสตราจารย์ ดร.เสมอ ถาน้อย)
รองอธิการบดี มหาวิทยาลัยพะเยา