การประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

รัตติกาล จอมประพันธ์ และ พาชิตชนัต ศิริพานิช***

บทคัดย่อ

งานวิจัยครั้งนี้เป็นการศึกษาวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยนำเสนอวิธีประมาณค่าสูญหายของตัวแปรตาม 4 วิธี คือ วิธีอัตราส่วนควอไทล์ที่ 1 วิธีอัตราส่วนควอไทล์ที่ 3 วิธีสมการถดถอย-อัตราส่วนควอไทล์ที่ 1 และวิธีสมการถดถอย-อัตราส่วนควอไทล์ที่ 3 ซึ่งพัฒนามาจากตัวประมาณอัตราส่วน พร้อมทั้งเปรียบเทียบประสิทธิภาพของวิธีที่นำเสนอ 4 วิธีดังกล่าวกับวิธีประมาณค่าสูญหายที่มีการใช้กันอยู่แล้ว 2 วิธี ได้แก่ วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ยและวิธีประมาณค่าสูญหายด้วยค่าการถดถอย โดยใช้ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์เป็นเกณฑ์ในการเปรียบเทียบ ภายใต้สถานการณ์ต่าง ๆ ซึ่งได้จากการจำลอง ผลการศึกษา พบว่า ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ของทุกวิธีมีแนวโน้มเพิ่มขึ้น เมื่อเปอร์เซ็นต์ของค่าสูญหายในตัวแปรตามเพิ่มขึ้น และเมื่อค่าความแปรปรวนของความคลาดเคลื่อนเพิ่มขึ้น นอกจากนั้น ในทุกสถานการณ์ วิธีสมการถดถอย-อัตราส่วนควอไทล์ที่ 1 วิธีสมการถดถอย-อัตราส่วนควอไทล์ที่ 3 และวิธีสมการถดถอย จะให้ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ที่ต่ำใกล้เคียงกัน (ความแตกต่างไม่มีนัยสำคัญทางสถิติที่ระดับ $\alpha = 0.05$) และยังเป็นวิธีการประมาณค่าที่มีประสิทธิภาพสูงในการประมาณค่าสูญหาย ส่วนวิธีค่าเฉลี่ย วิธีอัตราส่วนควอไทล์ที่ 1 และวิธีอัตราส่วนควอไทล์ที่ 3 นั้น ทั้ง 3 วิธีนี้เป็นวิธีที่มีประสิทธิภาพต่ำ และมีค่าความคลาดเคลื่อนสูง

คำสำคัญ: *ค่าสูญหาย การถดถอยพหุคูณ วิธีการประมาณค่าสูญหาย ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์*

---

* นักศึกษาปริญญาโท คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์
ถนนเสรีไทย แขวงคลองจั่น เขตบางกะปิ กรุงเทพฯ 10240
เมล: ruttikan_jomprapan@hotmail.com

** รองศาสตราจารย์ คณะบริหารธุรกิจ มหาวิทยาลัยธุรกิจบัณฑิตย์
110/1-4 ถนนประชาชื่น เขตหลักสี่ กรุงเทพฯ 10210
เมล: oui52@as.nida.ac.th

Missing Imputation in Multiple Linear Regression Analysis

Ruttikan Jomprapan and Pachitjanut Siripanich***

Abstract

The objective of this research is to propose imputation estimators when there are missing observations on dependent variable in the multiple linear regression analysis. The proposed ones are called ratio-Q1 (RQ1), ratio-Q3 (RQ3), regression-ratio-Q1 (RRQ1), and regression-ratio-Q3 (RRQ3). In various simulation situations, efficiency of the proposed estimators are compared to two existing methods, namely mean imputation and regression imputation, by using the mean absolute percentage error (MAPE) as a criterion. For each situation, linear regression model with 2 independent variables are considered under the assumption that the error is distributed as normal with various values of variances, sample sizes and percentages of missing observations on dependent variable. Findings reveal that the MAPE of all estimators increase as either the percentage of missing values or the variance of the error increases. Moreover, in all situations, RRQ1, RRQ3 and Regression imputation attain insignificantly different lowest values of MAPE. While the mean imputation, ratio-Q1 (RQ1), and ratio-Q3 (RQ3) are less efficient for their MAPE's are quite high.

Keywords: *Missing Data, Multiple Regression, Imputation, MAPE*

* Student, Master program in statistics, Graduate School of Applied Statistics, The National Institute of Development Administration (NIDA).
118 Sereethai Road, Klong-Chan, Bangkapi, Bangkok 10240, THAILAND.
E-mail: ruttikan_jomprapan@hotmail.com

** Associate Professor, Faculty of Business Administration, Dhurakij Pundit University
110/1-4 Prachachuen Road, Laksi, Bangkok 10210, THAILAND.
E-mail: oui52@as.nida.ac.th

1. บทนำ

การวิเคราะห์การถดถอยเป็นการวิเคราะห์ข้อมูลเชิงสถิติวิธีหนึ่งที่มีการประยุกต์ใช้ในด้านต่าง ๆ อย่างกว้างขวาง อาทิเช่น ด้านวิศวกรรม ด้านอุตสาหกรรม ด้านวิทยาศาสตร์ ด้านสังคมศาสตร์ และในด้านธุรกิจ เพื่อนำมาช่วยทำนายค่าตัวแปรตาม (Dependent Variable) หรือใช้การพยากรณ์สิ่งที่จะเกิดขึ้นในอนาคตโดยอาศัยความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรที่เป็นปัจจัยภายนอก ซึ่งในที่นี้จะเรียกว่า ตัวแปรอิสระ (Independent Variable) หากการวิเคราะห์การถดถอยมีตัวแปรอิสระเพียงตัวเดียวในการอธิบายตัวแปรตาม จะเรียกว่า การวิเคราะห์การถดถอยอย่างง่าย (Simple Regression) แต่หากมีตัวแปรอิสระมากกว่า 1 ตัวในการอธิบายตัวแปรตาม จะเรียกว่า การวิเคราะห์การถดถอยพหุคูณ (Multiple Regression) ปัญหาหนึ่งที่ผู้วิจัยพบบ่อย คือ ชุดข้อมูลที่รวบรวมมาได้นั้นมีข้อมูลไม่ครบถ้วนสมบูรณ์ ซึ่งอาจเกิดจากขั้นตอนการจัดเก็บข้อมูล การป้อนข้อมูล เครื่องมือจัดเก็บ การโอนถ่ายข้อมูลเกิดความผิดพลาด หรือเกิดจากข้อจำกัดของเทคโนโลยีที่ใช้งาน ทำให้เกิดข้อมูลสูญหาย ซึ่งส่งผลให้ไม่สามารถใช้ประโยชน์จากข้อมูลชุดนั้นได้เต็มที่

โดยปกติแล้วการวิเคราะห์การถดถอยพหุคูณ จะนำชุดข้อมูลที่ไม่มีค่าสูญหายในแต่ละตัวแปรมาใช้ในการพิจารณา แต่หากผู้วิจัยนำเอาชุดข้อมูลที่มีค่าสูญหายมาใช้ อาจทำให้เกิดปัญหาในการวิเคราะห์ ดังนั้น ผู้วิจัยบางคนอาจแก้ปัญหาโดยการตัดค่าสังเกตชุดนั้นทิ้งไป และวิเคราะห์ข้อมูลเท่าที่มีอยู่ ซึ่งอาจไม่เหมาะสม โดยเฉพาะอย่างยิ่ง กรณีที่ข้อมูลที่เหลืออยู่มีจำนวนน้อยกว่าที่วางแผนไว้มาก เป็นผลทำให้สูญเสียประสิทธิภาพทางสถิติไปได้มาก (ประชุม สุวัตถี, 2552: 518) ทำให้การประมาณค่าขาดความน่าเชื่อถือ ดังนั้น การจัดการปัญหาข้อมูลสูญหายจึงมีความสำคัญมาก

งานวิจัยโดยทั่วไปจะใช้การวิเคราะห์การถดถอยในการทำนายค่าของตัวแปรตาม ประมาณค่าเฉลี่ยหรือแนวโน้มของตัวแปรตาม โดยอาศัยอิทธิพลของตัวแปรอิสระที่มีผลต่อตัวแปรตามในการประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอย ซึ่งวิธีที่นิยมใช้ในการประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอย คือ วิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method – OLS) ซึ่งเป็นวิธีที่ให้ตัวประมาณที่มีคุณสมบัติที่ไม่เอนเอียงและมีความแปรปรวนต่ำสุดในบรรดาตัวประมาณเชิงเส้น แต่เมื่อค่าสังเกตสูญหายไปบางส่วนทำให้ประสิทธิภาพของตัวประมาณลดลงหรืออาจไม่เหมาะสมในการใช้วิธีกำลังสองน้อยที่สุด โดยปกติวิธีการแก้ปัญหาเมื่อมีค่าสังเกตสูญหาย อาจจะใช้วิธีตัดค่าสังเกตนั้นทิ้งไปดังที่ได้กล่าวไว้แล้วข้างต้น หรืออาจใช้เทคนิคเฉพาะสำหรับการวิเคราะห์ชุดข้อมูลที่มีค่าสังเกตสูญหายบางค่า หรือประมาณค่าสังเกตที่สูญหายด้วยวิธีการต่าง ๆ ก่อนที่จะวิเคราะห์ข้อมูลด้วยวิธีปกติโดยทำเสมือนมีข้อมูลสมบูรณ์

การจัดการกับข้อมูลสูญหายมีหลายวิธี ขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้น หากเลือกใช้วิธีจัดการกับข้อมูลสูญหายที่ไม่เหมาะสมย่อมส่งผลกระทบทำให้เกิดการบิดเบือนต่อผลการวิเคราะห์ได้ โดยรูปแบบของการสูญหาย สามารถแบ่งออกเป็น 3 ประเภท (Little และ Rubin,

1987) คือ

1. Missing completely at random (MCAR) เป็นลักษณะของข้อมูลสูญหายสำหรับตัวแปร Y เมื่อความน่าจะเป็นของค่าสูญหายของ Y ไม่มีความสัมพันธ์กับตัวมันเองและไม่มีความสัมพันธ์กับตัวแปรอื่น ๆ นั่นคือ ข้อมูลที่สูญหายเป็นอิสระกัน

2. Missing at random (MAR) เป็นลักษณะของข้อมูลสูญหายสำหรับตัวแปร Y เมื่อความน่าจะเป็นของค่าสูญหายของ Y ไม่มีความสัมพันธ์กับตัวมันเอง แต่อาจจะมีความสัมพันธ์หรือสามารถทำนายจากตัวแปรอื่น ๆ ได้

3. Not missing at random (NMAR) เป็นลักษณะของข้อมูลสูญหายสำหรับตัวแปร Y เมื่อความน่าจะเป็นของค่าสูญหายของ Y ไม่มีความสัมพันธ์กับค่าของตัวแปรอื่น ๆ แต่จะมีความสัมพันธ์กับตัวมันเอง

การวิจัยครั้งนี้ สนใจวิธีการจัดการกับปัญหาข้อมูลสูญหายโดยการประมาณค่าข้อมูลสูญหาย (Imputation) เมื่อลักษณะของข้อมูลสูญหายเป็นแบบ MAR และเนื้อหาที่นำเสนอในบทความนี้จะเริ่มจากวิธีการประมาณค่าสูญหาย 2 วิธีที่รู้จักกันโดยทั่วไป ตามด้วยการนำเสนอการประมาณค่าสูญหายซึ่งพัฒนามาจากการประมาณแบบอัตราส่วน (Ratio Estimation) พร้อมทั้งศึกษาคุณสมบัติของตัวประมาณที่นำเสนอ เปรียบเทียบกับวิธีการประมาณค่าสูญหายที่มีผู้นำเสนอไว้ สุดท้ายเป็นการอภิปรายผลการศึกษาและข้อเสนอแนะ

2. การประมาณค่าสูญหาย

ให้ $(y_1,x_{11},...,x_{k1})$, $(y_2,y_{12},...,x_{k2})$,...,$(y_n,x_{1n},...,x_{kn})$ เป็นค่าสังเกตจากตัวอย่างขนาด n โดยค่าสังเกตแต่ละค่าประกอบด้วยตัวแปร $k+1$ ตัว โดยไม่สูญเสียสภาพทั่วไป (Without Loss of Generality) สมมุติว่า $y_{r+1}, y_{r+2},...y_n$ คือ ค่าสูญหายที่ไม่สามารถสังเกตได้ ส่วนตัวแปร y อื่น ๆ และตัวแปร x ทั้งหมดสามารถสังเกตได้ (ไม่สูญหาย) สมมติว่าค่าสูญหายข้างต้นเป็นการสูญหายแบบ Missing at Random (MAR)

วิธีการจัดการกับข้อมูลสูญหายกรณีข้อมูลสูญหายเกิดจากการสูญหายบางตัวแปรของบางค่าสังเกต อาจใช้วิธีประมาณค่าสูญหาย (Imputation) ซึ่งแบ่งออกเป็นสองกลุ่มหลัก (Laaksonen, 2000) ได้แก่

1) Model-donor Imputation คือ การประมาณค่าที่ได้มาจากตัวแบบ (Model) ซึ่งมีอยู่หลากหลายวิธี เช่น Mean Imputation, Regression Imputation, Ratio Imputation และ Multiple Imputation

2) Real-donor Imputation คือ การประมาณค่าที่ได้จากเซตข้อมูลของค่าที่สังเกตได้ วิธีที่ใช้กันโดยทั่วไป เช่น Cold Deck Imputation, Hot Deck Imputation และ Nearest Neighbor Imputation

ในงานวิจัยครั้งนี้ ผู้วิจัยสนใจศึกษาวิธีการประมาณค่าสูญหายของตัวแปรตาม ซึ่งการประมาณค่าสูญหายอาจทำได้โดยใช้ตัวมันเองหรือใช้ตัวแปรช่วย (Auxiliary Variable) นั่นคือ ใช้ตัวแปร x (ตัวเดียวหรือหลายตัว) มาช่วยในการประมาณค่าตัวแปร y ที่สูญหาย ในที่นี้ วิธีประมาณค่าสูญหายที่ผู้วิจัยสนใจ คือ วิธีประมาณค่าแบบ Model-donor Imputation ซึ่งได้แก่ วิธีค่าเฉลี่ยและวิธีสมการถดถอย ซึ่งมีรายละเอียดของวิธีการประมาณค่า ดังนี้

2.1 วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย (Mean Imputation)

สมมุติให้ $y_1, y_2,...,y_r$ เป็นค่าสังเกต r ค่าจากทั้งหมด n ค่าและ $y_{r+1}, y_{r+2},...y_n$ เป็นค่าสูญหาย $n-r$ ค่าของ y ที่สังเกตไม่ได้ ในที่นี้ต้องการประมาณค่าสูญหายดังกล่าวด้วยค่าเฉลี่ย

วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย นำเสนอครั้งแรกโดย Wilks ในปี ค.ศ. 1932 ซึ่งเป็นการประมาณค่าของตัวแปรตามที่สูญหายโดยใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม ดังนี้

$$\hat{y}_M = \bar{y}^* = \frac{\sum_{i=1}^{r} y_i}{r^*} \tag{1}$$

เมื่อ $\bar{y}^*$ เป็นค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม r^* เป็นจำนวนข้อมูลที่ไม่สูญหายของตัวแปร y และ $i = 1, 2, 3, \ldots, r$

เนื่องจากวิธีนี้เป็นวิธีที่แทนค่าสูญหายด้วยค่าเฉลี่ยของข้อมูลที่สังเกตได้ (ไม่สูญหาย) ซึ่งค่าเฉลี่ยดังกล่าวของข้อมูลชุดหนึ่ง ๆ มีอยู่เพียงค่าเดียว ดังนั้น หากมีข้อมูลสูญหายมากกว่า 1 ค่า ก็จะประมาณค่าสูญหายทุกค่าด้วยค่าเฉลี่ยดังกล่าว ซึ่งจะส่งผลให้ค่าประมาณของข้อมูลที่สูญหายมีค่าเท่ากันหมด นั่นคือ ข้อมูลที่ประมาณค่าสูญหายแล้ว (Imputed Data) คือ $y_1, y_2,...,y_r, \bar{y},...,\bar{y}$ เห็นได้ชัดเจนว่า แม้จะมีข้อมูลสูญหายเพียงค่าเดียว หรือหลายค่าก็ตาม ความแปรปรวนของข้อมูล $y_1, y_2,...,y_r, \bar{y},...,\bar{y}$ ชุดนี้จะมีค่าน้อย ยิ่งจำนวนข้อมูลสูญหายมากขึ้น ความแปรปรวนก็จะน้อยลงด้วย ซึ่งส่งผลให้ค่าความคลาดเคลื่อนมาตรฐานของตัวประมาณของพารามิเตอร์ที่สนใจ เช่น ค่าความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง (Standard Error of the Sample Mean) มีค่าต่ำกว่าที่ควรจะเป็น (Underestimate)

2.2 วิธีประมาณค่าสูญหายด้วยการถดถอย (Regression Imputation)

วิธีนี้ใช้การประมาณค่าที่สูญหายโดยอาศัยความสัมพันธ์เชิงฟังก์ชันระหว่างตัวแปรตาม (y) กับตัวแปรอิสระ (x) กล่าวคือ พิจารณาสมการถดถอยของ y บน x ตัวหนึ่งหรือมากกว่าจากหน่วยตัวอย่างที่สังเกตค่า y และ x ได้ (r หน่วยแรกในตัวอย่าง) และประมาณค่า y ของหน่วยตัวอย่างที่สูญหาย ($n-r$ หน่วยหลังในตัวอย่าง) ด้วยค่าทำนายของ y ($\hat{y}$) โดยแทนค่าสังเกตของ x ที่สอดคล้องกับค่า y ที่สูญหายลงในสมการถดถอย (ประชุม สุวัตถี, 2554: 285) ในที่นี้ค่าของ x สามารถสังเกตได้ทุกค่า (ไม่มีค่า x สูญหาย) แนวความคิดดังกล่าวนี้เป็นผลจากการเสนอของ Buck ในปี 1960 ประกอบด้วย 3 ขั้นตอน คือ

ขั้นตอนที่ 1

วิเคราะห์การถดถอยเชิงเส้นโดยใช้ข้อมูล จากข้อสมมติข้างต้นที่ว่า $(y_1, x_{11},...,x_{k1})$, $(y_2, x_{12},...,x_{k2})$,..., $(y_r, x_{1r},...,x_{kr})$ เป็นค่าสังเกตที่เก็บรวบรวมได้ เพื่อประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด ซึ่งจะได้สมการถดถอย ดังนี้

$$\hat{y}^* = x^* \hat{\beta}^* \quad \text{โดยที่} \quad \hat{\beta}^* = \left(x^{*\prime} x^*\right)^{-1} x^{*\prime} y^* \tag{2}$$

เมื่อ y^* เป็นชุดข้อมูลจากค่าสังเกต r ค่าของ y ที่เก็บรวบรวมได้ และ x^* คือ ค่าสังเกต r ค่าของ x ซึ่งสอดคล้องกับ y^*

ขั้นตอนที่ 2

ประมาณค่าสูญหายของตัวแปร y_j เมื่อ $j = r+1, r+2, \ldots, n$ ด้วย $\hat{y}_j$ จากสมการถดถอยดังนี้

$$\hat{y}_j = \boldsymbol{x}_j \hat{\beta}^* \text{ เมื่อ } \boldsymbol{x}_j = (1, x_{1j}, x_{2j}, ..., x_{kj}) \tag{3}$$

ขั้นตอนที่ 3

ข้อมูลสมบูรณ์ของ y ($\tilde{y}$) ที่สามารถนำไปใช้ในการวิเคราะห์ข้อมูลโดยวิธีปกติ คือ ข้อมูลที่ได้จากการแทนค่าสูญหายด้วยค่าประมาณของข้อมูลสูญหาย (Imputation) ที่ได้จากสมการที่ 3 ดังนี้

$$\tilde{\boldsymbol{y}} = (y_1, y_2, \ldots, y_r, \hat{y}_{r+1}, \hat{y}_{r+2}, \ldots, \hat{y}_n)'$$

เมื่อ y_j คือ ค่าของ y ที่สังเกตได้ (ไม่สูญหาย), $j = 1, 2, \ldots, r$ และ $\hat{y}_j$ คือ ค่าประมาณของ y_j ที่สูญหาย โดยที่ $j = r+1, r+2, \ldots, n$.

ที่ผ่านมา มีนักวิจัยหลายท่านได้ทำการศึกษาเกี่ยวกับวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ยและวิธีประมาณค่าสูญหายด้วยการถดถอย เช่น ชุติมา ชัยมุสิก (2533) ศึกษาเปรียบเทียบการประมาณข้อมูลสูญหายในการวิเคราะห์การถดถอยเชิงพหุ พบว่า โดยส่วนใหญ่วิธีค่าเฉลี่ยให้ผลดีที่สุด ยกเว้นเมื่อมีขนาดตัวอย่างน้อยและจำนวนตัวแปรอิสระมาก วิธีประมาณค่าสูญหายด้วยการถดถอยจะให้ผลดีที่สุด ส่วนในงานวิจัยของวารุณี ตรีบำรุงศักดิ์ (2538) ได้ศึกษาการพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุเมื่อตัวแปรตามมีค่าสูญหาย พบว่า เมื่อความคลาดเคลื่อนสูง วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ยเป็นวิธีที่ดีในทุกสัดส่วนการสูญหายของตัวแปรตาม และสำหรับงานวิจัยของจริยา แสงสุวรรณ (2551) ได้ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ พบว่า เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีสมการถดถอยให้ค่าประมาณของ RMSE ลดลง นอกจากนี้ ยังมีงานวิจัยของ Christine Bono และคณะ (2007) ได้ศึกษาข้อมูลสูญหายใน The Center for Epidemiologic Studies Depression Scale ผลการศึกษา พบว่า วิธีค่าเฉลี่ยทุกวิธีมีความคล้ายคลึงกันกับวิธีประมาณค่าสูญหายด้วยค่าเฉลี่ยของวิธีสมบูรณ์ ยกเว้นวิธีประมาณค่าสูญหายด้วยการถดถอย ในการแทนค่าสูญหายไม่ทำให้ระดับนัยสำคัญของข้อสรุปเปลี่ยนแปลงไป

นอกจากวิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย และวิธีประมาณค่าสูญหายด้วยการถดถอยที่ผู้วิจัยได้กล่าวถึงไปแล้วนั้นยังมีวิธีประมาณค่าสูญหายอีกหลายวิธีที่ผู้วิจัยไม่ได้กล่าวถึง เนื่องจากอยู่นอกขอบเขตความสนใจของผู้วิจัย

3. วิธีประมาณค่าสูญหายที่นำเสนอ

3.1 ทฤษฎีพื้นฐาน

การประมาณค่าเฉลี่ย (μ_y) ของตัวแปร y ใด ๆ ด้วยตัวประมาณอัตราส่วน (Ratio Estimator) เป็นวิธีที่รู้จักกันอย่างกว้างขวาง และเป็นตัวประมาณที่อาศัยความสัมพันธ์ระหว่างตัวแปรช่วย (Auxiliary Variable) X กับตัวแปรตาม Y (Cochran, 1977: 150-151) และตัวประมาณอัตราส่วนของค่าเฉลี่ยของ y มีสูตรการคำนวณ ดังนี้ : $\hat{\mu}_Y = \bar{y}\left(\frac{\mu_X}{\bar{x}}\right)$ เมื่อ $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ และ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ คือ ค่าเฉลี่ยตัวอย่างของตัวแปรช่วย X และตัวแปรตาม Y ตามลำดับ และ μ_x คือ ค่าเฉลี่ยของตัวแปรช่วย X ซึ่งทราบค่า นอกจากนี้ ยังพบว่า ความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของ $\hat{\mu}_y$ คือ $MSE(\hat{\mu}_Y) \cong \mu_Y^2 \frac{f}{n}(C_y^2 + C_x^2 - 2\rho C_y C_x)$ โดยที่ $f = \frac{N-n}{N}$, $C_y^2 = \frac{S_y^2}{\mu_Y^2}$, $C_x^2 = \frac{S_x^2}{\mu_X^2}$, $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu_X)^2$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu_Y)^2$, N คือ ขนาดประชากร n คือ ขนาดตัวอย่าง และ ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร Y และ X

ในกรณีที่มีตัวแปรช่วย 2 ตัว ตัวประมาณอัตราส่วนของค่าเฉลี่ยของ y คำนวณได้จากสูตร $\hat{\mu}_Y = \bar{y}\left(\frac{\mu_{x_1}}{\bar{x}_1}\right)\left(\frac{\mu_{x_2}}{\bar{x}_2}\right)$ และความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\mu}_y$ คือ $MSE\left(\hat{\mu}_Y\right) \cong \mu_Y^2 \frac{f}{n}(C_y^2 + C_{x_1}^2 + C_{x_2}^2 - 2\rho_{yx_1}C_yC_{x_1} - 2\rho_{yx_2}C_yC_{x_2} + 2\rho_{x_1x_2}C_{x_1}C_{x_2})$ เมื่อ $f = \frac{N-n}{N}$, $C_y^2 = \frac{S_y^2}{\mu_Y^2}$, $C_{x_k}^2 = \frac{S_{x_k}^2}{\mu_{X_k}^2}$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu_Y)^2$, $S_{x_k}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_{ki} - \mu_{X_k})^2$, $S_{yx_k} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu_Y)(X_{ki} - \mu_{X_k})^2$ และ $\rho_{yx_k} = \frac{S_{yx_k}}{\sqrt{S_y^2}\sqrt{S_{x_k}^2}}$ โดยที่ N คือ ขนาดของประชากร n คือ ขนาดของตัวอย่าง และ k = 1, 2 (Abu-Dayyeh และคณะ, 2003: 287-298)

นอกจากนั้นแล้วยังมีผู้คิดค้นและพัฒนาตัวประมาณอัตราส่วนของ μ_y โดยนำสารสนเทศของตัวแปรช่วยมาใช้ประกอบในการประมาณค่าด้วยวิธีต่าง ๆ เช่น Sisodia และ Dwivedi (1981: 13-18) ได้นำสัมประสิทธิ์ความแปรผันของตัวแปร $X(C_x)$ เข้ามาใช้ ทำให้ได้ตัวประมาณ คือ $\bar{y}_{SD} = \bar{y}\left(\frac{\mu_x + C_x}{\bar{x} + C_x}\right)$ ต่อมา Singh และ Kakran (1993: 894) ได้ปรับปรุงตัวประมาณ $\bar{y}_{SD}$ โดยการแทนที่ C_x ด้วยสัมประสิทธิ์ความโด่งของประชากร (Coefficient of Kurtosis: $\beta_2(x)$) ซึ่งจะได้ตัวประมาณคือ $\bar{y}_{SK} = \bar{y}\left(\frac{\mu_x + \beta_2(x)}{\bar{x} + \beta_2(x)}\right)$ นอกจากนั้น ยังมีตัวประมาณของ Singh and Tailor (2003: 555-560) ที่ได้นำสัมประสิทธิ์สหสัมพันธ์ (ρ) มาใช้ คือ $\bar{y}_{ST} = \bar{y}\left(\frac{\mu_x + \rho}{\bar{x} + \rho}\right)$ และยังมีตัวประมาณของ Al-Omari และคณะ (2009: 97-108) ที่นำประโยชน์จากค่าควอไทล์ที่ 1 และ 3 ของตัวแปร X มาประยุกต์ใช้ในตัวประมาณแบบอัตราส่วน ซึ่งจะได้ตัวประมาณ

$$\bar{y}_{Al1} = \bar{y}\left(\frac{\mu_x + q_1}{\bar{x} + q_1}\right) \text{ หรือ } \bar{y}_{Al3} = \bar{y}\left(\frac{\mu_x + q_3}{\bar{x} + q_3}\right) \quad (4)$$

Ray และ Singh (1981: 147-151) เป็นผู้หนึ่งที่นำค่าทำนายของตัวแปร Y เมื่อกำหนดค่าของตัวแปร X มาประกอบในการประมาณค่าเฉลี่ยของประชากร (โดยการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด) ทำให้ได้ตัวประมาณแบบอัตราส่วน คือ $\bar{y}_{RS} = \frac{\bar{y} + b_{LS}(\bar{x}^{\alpha} - \mu_x^{\alpha})}{\bar{x}^{\gamma}}\mu_x^{\gamma}$ ต่อมา Kadilar และ Cingi (2004: 893-902) ได้นำตัวประมาณ $\bar{y}_{RS}$ มาปรับปรุงโดยกำหนดให้ค่าคงที่ α และ γ เท่ากับ 1 ซึ่งทำให้ได้ตัวประมาณแบบอัตราส่วน คือ $\bar{y}_{KC} = \frac{\bar{y} + b_{LS}(\mu_x - \bar{x})}{\bar{x}}\mu_x$

โดยที่ $b_{LS} = \frac{s_{xy}}{s_x^2}$ คือ ตัวประมาณกำลังสองน้อยที่สุดของสัมประสิทธิ์การถดถอยในตัวแบบเชิงเส้นอย่างง่ายระหว่างตัวแปร Y และตัวแปร X, s_x^2 คือ ความแปรปรวนของตัวอย่างของตัวแปร X, s_{xy} คือ ความแปรปรวนร่วมของตัวอย่างของตัวแปร Y และตัวแปร X

ปัญหาสำคัญประการหนึ่งในการวิเคราะห์ข้อมูลเพื่อประมาณค่าเฉลี่ย นั่นคือ ข้อมูลเกิดการสูญหาย ดังนั้น ในงานวิจัยของ Nuanpan Nangsue (2009) ซึ่งได้นำตัวประมาณอัตราส่วนและตัวประมาณการถดถอยมาใช้ในประมาณค่าเฉลี่ยเมื่อข้อมูลเกิดการสูญหาย ดังนี้

$$\bar{y}_{R1} = \bar{y}_{s1s3}\left(\frac{\mu_X}{\bar{x}_{s1s2}}\right), \quad \bar{y}_{R2} = \bar{y}_{s1s3} + b_1(\mu_X - \bar{x}_{s1s2}), \quad \bar{y}_{R3} = \bar{y}_{s1s3} + b_1(\bar{x}_{s1s2} - \bar{x}_{s1}),$$

$$\bar{y}_{R4} = \bar{y}_{s1s3}\left(\frac{\mu_X}{\bar{x}_{s1s2}}\right)^{b_1}, \quad \bar{y}_{R5} = \bar{y}_{R4} + b_1(\mu_X - \bar{x}_{s1s2}) \text{ และ} \quad \bar{y}_{R6} = \bar{y}_{R2}\left(\frac{\mu_X}{\bar{x}_{s1s2}}\right)^{b_1}$$

โดยที่ s_1 คือ ข้อมูลของตัวอย่างตัวแปร x และ y สมบูรณ์ s_2 คือ ข้อมูลของตัวอย่างตัวแปร x สมบูรณ์ แต่ตัวแปร y มีค่าสูญหาย s_3 คือ ข้อมูลของตัวอย่างตัวแปร y สมบูรณ์ แต่ตัวแปร x มีค่าสูญหาย $\hat{\beta} = b_1 = \frac{rS_y}{S_x}$ คือ ค่าสัมประสิทธิ์การถดถอย

3.2 ตัวประมาณค่าสูญหายที่นำเสนอ

เนื่องจากตัวประมาณอัตราส่วนที่ได้กล่าวมา เป็นตัวประมาณที่อาศัยความสัมพันธ์ระหว่างตัวแปรช่วย X กับตัวแปรตาม Y ดังนั้น ผู้วิจัยจึงสนใจนำตัวประมาณอัตราส่วนมาประยุกต์เพื่อใช้ในการประมาณค่าสูญหายของตัวแปรตาม ซึ่งตัวประมาณอัตราส่วนที่ผู้วิจัยสนใจศึกษา คือ ตัวประมาณอัตราส่วน (สมการที่ (4)) ของ Al-Omari และคณะ (2009: 97-108) เป็นตัวประมาณค่าเฉลี่ยของประชากรโดยอาศัยควอไทล์ที่ 1 และ 3 ของตัวแปรตาม อย่างไรก็ตาม ตัวประมาณนี้ใช้สำหรับข้อมูลสมบูรณ์และมีเงื่อนไขว่า ทราบค่าเฉลี่ยที่แท้จริงของตัวแปร X (ตัวแปรช่วย) ในกรณีที่ศึกษานี้เป็นกรณีที่มีข้อมูลสูญหายและโดยทั่วไปในการวิเคราะห์การถดถอย เราจะไม่ทราบค่าเฉลี่ยที่แท้จริงของตัวแปรช่วย (X) ดังนั้น จากตัวประมาณของ Al-Omari และคณะ ตามสมการที่ (4) ผู้วิจัยได้ประมาณค่า $\bar{x}_{in}$ มาแทนค่า μ_x โดยกำหนดให้ $\bar{x}_{in}$ แทนค่าเฉลี่ยตัวอย่างที่สมบูรณ์ของตัวแปร x และ $\bar{x}_{ir}$ แทนค่าเฉลี่ยตัวอย่างที่ตัดค่าสูญหายออกของตัวแปร x ทำให้ได้ตัวประมาณค่าข้อมูลสูญหาย 2 ตัว ในที่นี้จะเรียกว่า RQ1 และ RQ3 ซึ่งมีสูตร ดังนี้

$$\hat{y}_{RQ1} = \hat{y}_j = \bar{y}\prod_{i=1}^{k}\left(\frac{\bar{x}_{in} + q_1}{\bar{x}_{ir} + q_1}\right) \quad \text{และ} \quad \hat{y}_{RQ3} = \hat{y}_j = \bar{y}\prod_{i=1}^{k}\left(\frac{\bar{x}_{in} + q_3}{\bar{x}_{ir} + q_3}\right) \tag{5}$$

อย่างไรก็ตาม การใช้ตัวประมาณเพียงค่าเดียวแทนค่าสูญหายทุกตัว อาจไม่เหมาะสมเนื่องจากการใช้ค่าคงที่แทนค่าสูญหายทุกตัวทำให้ความแปรปรวนของตัวประมาณมีค่าต่ำกว่าที่ควรจะเป็น (Underestimate) ดังนั้น ผู้วิจัยจึงนำค่าประมาณจากวิธีสมการถดถอย คือ $\hat{y}_{reg}$ มาแทน $\bar{y}$ ในสมการ (5) โดยการแทนค่า $\bar{y}$ ด้วยค่าประมาณจากวิธีสมการถดถอย ($\hat{y}_{reg}$) นั้น เนื่องจากค่าเฉลี่ยเป็นค่ากลางหรือที่เรียกว่า การวัดแนวโน้มเข้าสู่ส่วนกลางของข้อมูลซึ่งเป็นการนำค่าของข้อมูลทุกค่ามาคำนวณหาค่าเฉลี่ย ด้วยสาเหตุนี้ค่าเฉลี่ยที่ได้จากการคำนวณอาจสูงหรือต่ำมากเกินไปเนื่องจากอิทธิพลของค่าสังเกตบางค่าที่สูงหรือต่ำกว่าปกติ (Outlier) จึงทำให้เกิดความคลาดเคลื่อนของการประมาณค่าสูง ดังนั้น ผู้วิจัยจึงใช้ค่าประมาณจากวิธีสมการถดถอยมาแทน เนื่องจากเป็นค่าที่ต้องอาศัยความสัมพันธ์ระหว่างตัวแปรช่วยที่ถูกกำหนดขึ้นกับตัวแปรตาม ซึ่งจะเรียกว่า RRQ1 และ RRQ3 ดังนี้

$$\hat{y}_{RRQ1} = \hat{y}_j = \hat{y}_{reg}\prod_{i=1}^{k}\left(\frac{\bar{x}_{in}+q_1}{\bar{x}_{ir}+q_1}\right) \text{ และ } \hat{y}_{RRQ3} = \hat{y}_j = \hat{y}_{reg}\prod_{i=1}^{k}\left(\frac{\bar{x}_{in}+q_3}{\bar{x}_{ir}+q_3}\right) \quad (6)$$

เมื่อ $\hat{y}_j$ คือ ค่าประมาณของค่าสูญหายตัวที่ j, $\bar{y}$ คือ ค่าเฉลี่ยตัวแปร y ที่สมบูรณ์, q_1 และ q_3 คือ ค่าควอไทล์ที่ 1 และ 3 ของตัวแปร x ขนาด n ตามลำดับ k คือ จำนวนตัวแปร x (k = 2), i = 1, 2, 3, ..., r (ค่าที่สมบูรณ์) และ j= r+1, r+2, ..., n (ค่าที่สูญหาย)

4. การจำลอง

4.1 ขั้นตอนการดำเนินงาน

ในการวิจัยครั้งนี้ ได้ทำการจำลองข้อมูลในการวิจัยตามสถานการณ์ต่าง ๆ ด้วยโปรแกรม SAS เพื่อสร้างข้อมูลให้เป็นไปตามการวิจัยโดยจะกระทำซ้ำกัน 1,000 ครั้ง ในแต่ละสถานการณ์ ตามรายละเอียด ขั้นตอนการดำเนินงาน ดังนี้

1. สร้างข้อมูลประชากรของตัวแปรอิสระ X_1, X_2 และข้อมูลของความคลาดเคลื่อน (ε) แต่ละประชากรมีขนาด 100,000 หน่วย และมีการแจกแจงปกติ ซึ่งมีฟังก์ชันความหนาแน่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty$$

เมื่อ ค่าคาดหวัง คือ $E(X) = \mu$ และความแปรปรวน คือ $V(X) = \sigma^2$ ซึ่งในงานวิจัยครั้งนี้กำหนดให้ตัวแปรอิสระ $X_1 \sim N(\mu = 3, \sigma^2 = 2.25)$ ตัวแปรอิสระ $X_1 \sim N(\mu = 5, \sigma^2 = 4)$ และความคลาดเคลื่อน $\varepsilon \sim N(\mu = 0, \sigma_\varepsilon^2)$ โดยที่ σ_ε^2 = 0.5, 1.0, 1.5 และ 2.0

2. สุ่มตัวอย่างขนาด n จากแต่ละประชากรของตัวแปรอิสระ X_1, X_2 และความคลาดเคลื่อน (ε) โดยใช้วิธีการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) ขนาดตัวอย่าง (n) ในการศึกษาครั้งนี้ คือ 20, 40, 60 และ 100 ในการสุ่มตัวอย่างแต่ละสถานการณ์จะกระทำซ้ำเป็นจำนวน 1,000 รอบ

3. กำหนดค่าสัมประสิทธิ์การถดถอยในตัวแบบเป็นดังนี้: $\beta_0 = 0.5$, $\beta_1 = 1$ และ $\beta_2 = -0.3$

4. สร้างตัวแปรตาม y_i; $i = 1, 2, \ldots n$, ตามตัวแบบการถดถอยเชิงเส้น ดังนี้

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

โดยตัวแปรอิสระ (x_{1i}, x_{2i}) สัมประสิทธิ์การถดถอย (β_0, β_1, β_2) และค่าความคลาดเคลื่อน (ε_i) เป็นไปตามที่กำหนดในขั้น 1-3 ข้างต้น

5. คำนวณหาจำนวนข้อมูลที่สูญหายและสุ่มตำแหน่งที่สูญหายของข้อมูลตัวแปรตาม

5.1 คำนวณหาจำนวนข้อมูลที่สูญหายจากสูตร

$$\textit{จำนวนข้อมูลที่สูญหาย} = \left\lceil \frac{np_{missing}}{100} \right\rceil$$

เมื่อ n คือ ขนาดตัวอย่าง $p_{missing}$ คือ ร้อยละของข้อมูลที่สูญหาย ในที่นี้กำหนดให้เป็น 10%, 15%, 20% และ $\lceil a \rceil$ คือ จำนวนเต็มที่น้อยที่สุด ที่มีค่ามากกว่าหรือเท่ากับ a

5.2 สุ่มตำแหน่งที่สูญหายของตัวแปรตาม

6. ประมาณค่าสูญหายของข้อมูลตัวแปรตามทั้ง 6 วิธี คือ

6.1 วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย (Mean Imputation – MI) ตามสูตรในสมการที่ (1)

6.2 วิธีประมาณค่าสูญหายด้วยการถดถอย (Regression Imputation – RI) ตามสูตรในสมการที่ (3)

6.3 วิธีอัตราส่วนควอไทล์ที่ 1 (RQ1) ตามสูตรในสมการที่ (5)

6.4 วิธีอัตราส่วนควอไทล์ที่ 3 (RQ3) ตามสูตรในสมการที่ (5)

6.5 วิธีสมการถดถอย – อัตราส่วน ควอไทล์ที่ 1 (RRQ1) ตามสูตรในสมการที่ (6)

6.6 วิธีสมการถดถอย – อัตราส่วน ควอไทล์ที่ 3 (RRQ3) ตามสูตรในสมการที่ (6)

หมายเหตุ การคำนวณค่าสูญหายสำหรับวิธี RQ1, RQ3, RRQ1 และ RRQ3 นั้น ใช้ค่าของตัวแปรอิสระ (X_1 และ X_2) เป็นตัวแปรช่วย (Auxiliary Variable)

7. จากข้อมูลที่แทนค่าสูญหายจากวิธีการประมาณค่าสูญหายวิธีต่าง ๆ ทั้ง 6 วิธีในขั้นตอนที่ 6 แล้ว นำไปประมาณค่าสัมประสิทธิ์การถดถอยใหม่ด้วยวิธีกำลังสองน้อยที่สุด เพื่อหาสมการ

ถดถอยเชิงเส้นพหุในการพยากรณ์

8. เปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามจากการประมาณค่าด้วยเกณฑ์วิธีค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error: MAPE)

9. สรุปผลการศึกษา

4.2 เกณฑ์ที่ใช้ในการตัดสินใจ

สมมุติให้ $\tilde{\boldsymbol{y}} = (y_1, y_2, \ldots, y_r, \tilde{y}^*_{r+1}, \tilde{y}^*_{r+2}, \ldots, \tilde{y}^*_n)'$ เป็นเวคเตอร์ของข้อมูลสมบูรณ์ของ y โดยที่ y_j, $j = 1, 2, \ldots, r$ คือ ค่าของ y ที่สังเกตได้ (ไม่สูญหาย) และ $\tilde{y}^*_j$ คือ ค่าประมาณของ y_j ที่สูญหาย ด้วยวิธีใดวิธีหนึ่ง เมื่อ $j = r+1, r+2, \ldots, n$

ให้ $\hat{\tilde{\boldsymbol{y}}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_r, \hat{\tilde{y}}^*_{r+1}, \hat{\tilde{y}}^*_{r+2}, \ldots, \hat{\tilde{y}}^*_n)'$ เป็นเวคเตอร์ของค่าทำนายของที่ได้จากการสารูป (fit) ข้อมูล $\tilde{\boldsymbol{y}}$ กับ $\boldsymbol{X}$ ภายใต้ตัวแบบวิเคราะห์การถดถอยเชิงเส้น โดยที่

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}_{n\times 3} \quad \text{และ} \quad \hat{\tilde{\boldsymbol{y}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}$$

ในรอบที่ t ของการจำลอง ค่าเฉลี่ยเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error) ซึ่งแทนด้วยสัญลักษณ์ $\overline{MAPE}_t$ คำนวณได้จากสูตร

$$\overline{MAPE}_t = \frac{\sum_{j=1}^{r}\left|\frac{(y_j - \hat{y}_j)}{y_j}\right| + \sum_{j=r+1}^{n}\left|\frac{(\tilde{y}^*_j - \hat{\tilde{y}}^*_j)}{\tilde{y}^*_j}\right|}{n}\times 100 \qquad (7)$$

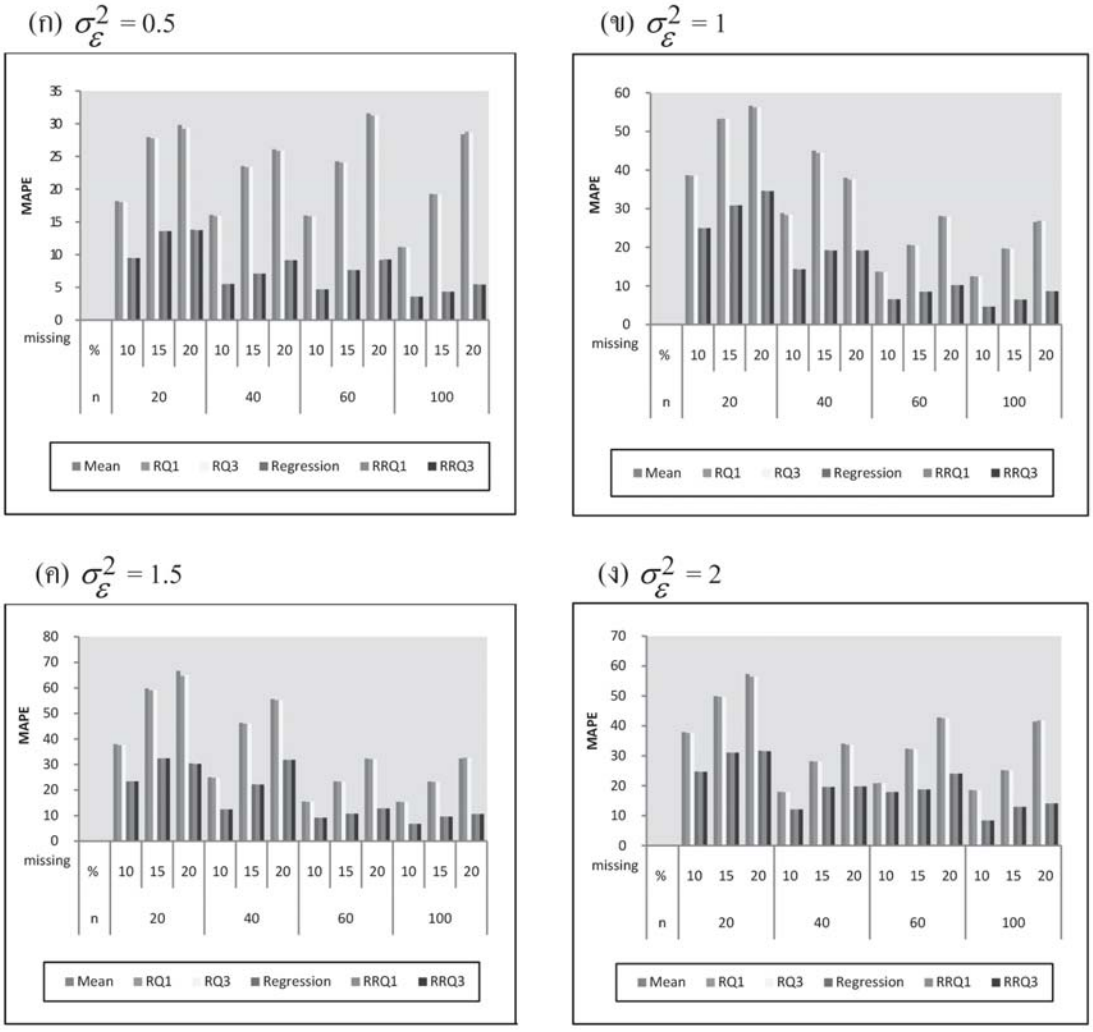
เกณฑ์ที่ใช้ในการตัดสินประสิทธิภาพของวิธีประมาณค่าการสูญหาย คือ $\overline{\overline{MAPE}}$ ซึ่งได้แก่ ค่าเฉลี่ย 1,000 รอบของ $\overline{MAPE}_t$ และคำนวณได้ดังนี้

$$\overline{\overline{MAPE}} = \frac{\sum_{t=1}^{1,000}\overline{MAPE}_t}{1,000}$$

เห็นได้ว่า ถ้าวิธีประมาณค่าสูญหายวิธีใดทำให้ $\overline{\overline{MAPE}}$ มีค่าต่ำกว่าวิธีอื่น นั่นแสดงว่าเป็นวิธีประมาณค่าสูญหายวิธีดังกล่าว เป็นวิธีที่ดีกว่าวิธีอื่น

5. ผลการศึกษา

ผลการศึกษาในภาพรวม พบว่า ค่า $\overline{\overline{MAPE}}$ ของทั้ง 4 วิธีที่นำเสนอ (วิธี RQ1, RQ3, RRQ1 และ RRQ3) มีค่าสูงขึ้นเมื่อสัดส่วนของค่าสูญหายเพิ่มขึ้น และค่า $\overline{\overline{MAPE}}$ ของทุกวิธีมีการกระจายมากขึ้นเมื่อความแปรปรวนของความคลาดเคลื่อน (ε) มีค่าสูงขึ้น ดังเห็นได้จากผลการศึกษาซึ่งแสดงไว้ในตารางที่ ก.1-ก.4 ในภาคผนวก และสรุปเป็นภาพรวมในภาพที่ 1



ภาพที่ 1: ค่า MAPE ของวิธีการประมาณค่าสูญหาย จำแนกตามค่าความแปรปรวนของความคลาดเคลื่อน (σ_ε^2)

เมื่อพิจารณาในรายละเอียดแต่ละวิธี พบว่า $\overline{\overline{MAPE}}$ ของวิธี RRQ1 และ RRQ3 มีค่าต่างกันน้อยมากทุกกรณี ในทำนองเดียวกัน $\overline{\overline{MAPE}}$ ของวิธี RQ1 และ RQ3 ก็มีค่าต่างกันน้อยมากทุกกรณี เมื่อเปรียบเทียบ $\overline{\overline{MAPE}}$ ของวิธีประมาณค่าสูญหายทั้ง 2 คู่นี้ พบว่า $\overline{\overline{MAPE}}$ ของวิธี RRQ1 และ RRQ3 มีค่าน้อยกว่า $\overline{\overline{MAPE}}$ ของวิธี RQ1 และ RQ3 อย่างชัดเจน นอกจากนี้ ยังพบว่าเมื่อขนาดตัวอย่างไม่สูงนัก (n = 20) $\overline{\overline{MAPE}}$ ของวิธี RQ1 และ RQ3 มีค่าประมาณ 1.5-2.2 เท่าของวิธี RRQ1 และ RRQ3 โดยที่สัดส่วน $\frac{\overline{\overline{MAPE(RQ1)}}}{\overline{\overline{MAPE(RRQ1)}}}$ และ $\frac{\overline{\overline{MAPE(RQ3)}}}{\overline{\overline{MAPE(RRQ3)}}}$ จะมีค่าสูงขึ้นเล็กน้อยเมื่อสัดส่วนของข้อมูลสูญหายมีค่าเพิ่มขึ้น แต่เมื่อตัวอย่างมีขนาดเพิ่มขึ้น (n = 40, 60, 100) พบว่า สัดส่วนของ $\overline{\overline{MAPE}}$ ดังกล่าวมีแนวโน้มที่จะมีค่าลดลงเมื่อความแปรปรวนของความคลาดเคลื่อน (σ_ε^2) มีค่าลดลง และสัดส่วนของ $\overline{\overline{MAPE}}$ ดังกล่าวมีแนวโน้มที่จะมีค่าเพิ่มขึ้นเมื่อตัวอย่างมีขนาดเพิ่มขึ้น หรือสัดส่วนของข้อมูลสูญหายมีค่าเพิ่มขึ้น

เมื่อเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุด้วยวิธีต่าง ๆ 4 วิธีที่นำเสนอ (วิธี RQ1, RQ3, RRQ1 และ RRQ3) กับอีก 2 วิธีเดิม ๆ ได้แก่ วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย (Mean Imputation – MI) และวิธีประมาณค่าสูญหายด้วยการถดถอย (Regression Imputation – RI) โดยใช้ค่า MAPE เป็นเกณฑ์ในการเปรียบเทียบ พบว่า $\overline{\overline{MAPE}}$ ของวิธี MI มีค่ามากกว่าวิธี RQ1, RQ3 ทุกกรณี ยกเว้น กรณีที่ความแปรปรวนของความคลาดเคลื่อนมีค่าน้อย (σ_ε^2 = 0.5) สัดส่วนของข้อมูลสูญหายมีค่ามาก (20%) และตัวอย่างมีขนาดใหญ่ (n = 60, 100) อย่างไรก็ตาม $\overline{\overline{MAPE}}$ ของวิธี RQ1, RQ3 และ MI ก็มีค่าแตกต่างกันเพียงเล็กน้อยเท่านั้น ในทำนองเดียวกัน ในทุกกรณี $\overline{\overline{MAPE}}$ ของวิธี RI กับวิธี RRQ1, RRQ3 มีค่าแตกต่างกันเพียงเล็กน้อยเท่านั้น

6. สรุปและอภิปราย

การศึกษาครั้งนี้นำเสนอวิธีการประมาณค่าสูญหายซึ่งพัฒนามาจากวิธีประมาณค่าแบบสัดส่วน วิธีที่นำเสนอมีทั้งสิ้น 4 วิธี ได้แก่ วิธี RQ1, RQ3, RRQ1 และ RRQ3 โดย 2 วิธีแรกเป็นวิธีประมาณค่าสูญหายทุกค่าด้วยค่าเดียวกัน (เท่ากัน) หมด ดังสูตรการคำนวณในสมการ (5) ส่วน 2 วิธีหลัง เป็นวิธีประมาณค่าสูญหายแต่ละค่าโดยใช้ตัวแปรช่วย (Auxiliary Variable) ที่สอดคล้องกับตัวแปรสูญหายแต่ละตัว ดังสูตรการคำนวณในสมการ (6) ผลการศึกษา พบว่าวิธีการประมาณค่าสูญหาย RRQ1 และ RRQ3 มีประสิทธิภาพดีกว่าวิธี RQ1 และ RQ3 อย่างชัดเจน และเมื่อเปรียบเทียบกับวิธีที่มีผู้นำเสนอไว้แล้ว 2 วิธี คือ วิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ย (Mean Imputation – MI) และวิธีการประมาณค่าสูญหายด้วยการถดถอย (Regression Imputation – RI) พบว่า วิธี RI มีประสิทธิภาพดีใกล้เคียงกับวิธี RRQ1 และ RRQ3 ส่วนวิธี MI ก็มีประสิทธิภาพพอ ๆ กับวิธี RQ1 และ RQ3 ซึ่งมีประสิทธิภาพด้อยกว่า 3 วิธีแรก ทั้งนี้

ตัววัดประสิทธิภาพที่กล่าวถึง ได้แก่ ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error: MAPE)

เมื่อวัดประสิทธิภาพของวิธีประมาณค่าสูญหายด้วยค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (MAPE) กล่าวคือ วิธีประมาณค่าสูญหายมีประสิทธิภาพสูงถ้า MAPE มีค่าต่ำ ผลการศึกษาชี้ให้เห็นว่า วิธีประมาณค่าสูญหายที่ใช้รูปแบบค่าคงที่ในลักษณะของการใช้ค่าเฉลี่ยในการประมาณค่าสูญหาย ดังเช่นวิธี Mean, RQ1 และRQ3 นั้น ทำให้ประสิทธิภาพมีค่าต่ำ แต่เมื่อประมาณค่าสูญหายโดยอาศัยความสัมพันธ์เชิงเส้นระหว่างตัวแปรช่วย (Auxiliary Variable) กับตัวแปรตามที่กำหนด ทำให้ประสิทธิภาพมีค่าสูง ดังเช่นวิธี RRQ1, RRQ3 และRI

อย่างไรก็ตาม วิธีประมาณค่าสูญหายทุกวิธี จะมีประสิทธิภาพต่ำลงเมื่อสัดส่วนของค่าสูญหายเพิ่มขึ้น หรือความแปรปรวนของความคลาดเคลื่อน (σ_ε^2) มีค่าสูงขึ้น และวิธีประมาณค่าสูญหายทุกวิธี จะมีประสิทธิภาพสูงขึ้นเมื่อตัวอย่างมีขนาดสูงขึ้น

ผลงานวิจัยของ Al-Omari และคณะ (2009) พบว่า การนำค่าควอไทล์ที่ 1 มาใช้จะให้ประสิทธิภาพที่ดีกว่าการใช้ค่าควอไทล์ที่ 3

นอกจากนั้น เมื่อพิจารณาผลการวิจัยในวิธีที่ผู้วิจัยได้นำเสนอนั้น คือ วิธี RQ1, วิธี RQ3, วิธี RRQ1 และวิธี RRQ3 โดยส่วนใหญ่เมื่อพิจารณาจากค่า MAPE แล้ววิธีที่ใช้ค่าของควอไทล์ที่ 1 จะมีประสิทธิภาพมากกว่าการใช้ค่าของควอไทล์ที่ 3 ซึ่งสอดคล้องกับผลงานวิจัยของ Al-Omari และคณะ (2009) ที่กล่าวว่าผลการวิจัยที่นำค่าควอไทล์ที่ 1 มาใช้จะให้ประสิทธิภาพที่ดีกว่าการใช้ค่าควอไทล์ที่ 3 อีกทั้งผลที่ใช้วิธีในรูปแบบที่ต้องอาศัยความสัมพันธ์เชิงเส้นระหว่างตัวแปรช่วยกับตัวแปรตาม จะมีประสิทธิภาพสูง ซึ่งสอดคล้องกับ Little and Rubin (1987) ที่กล่าวว่าการนำค่าของตัวแปร โดยที่ i = 1, 2, ..., r มาอธิบายค่าของตัวแปร y เป็นเทคนิคที่ดี ดังนั้น หากตัวแปรช่วยกับตัวแปรตามมีความสัมพันธ์เชิงเส้น ควรเลือกใช้วิธีประมาณค่าสูญหาย วิธี RRQ1, วิธี RRQ3 หรือวิธี Regression จะเหมาะสมกว่าการประมาณค่าสูญหายด้วยค่าคงที่ (เช่นวิธี MI, วิธี RQ1 หรือวิธี RQ3) เนื่องจากเป็นวิธีที่ทำให้มีประสิทธิภาพสูง

7. ข้อเสนอแนะ

7.1 ด้านการนำไปใช้ประโยชน์

การศึกษาวิจัยครั้งนี้ พบว่า การประมาณค่าสูญหายด้วยค่าเพียงค่าเดียว ไม่ว่าจะเป็นวิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย (Mean Imputation) หรือวิธีที่นำเสนอซึ่งได้แก่ วิธี RQ1 และวิธี RQ3 ก็ตาม วิธีดังกล่าวทั้ง 3 วิธีมีประสิทธิภาพด้อยกว่าวิธีการประมาณค่าสูญหายที่นำเสนออีก 2 วิธี ได้แก่ วิธี RRQ1 และ RRQ3 ส่วนวิธีประมาณค่าสูญหายด้วยการถดถอย (Regression imputation) แม้จะมีประสิทธิภาพพอ ๆ กับวิธี RRQ1 และวิธี RRQ3 ดังได้กล่าวแล้วในการ

อภิปรายผล แต่หากพิจารณาให้ละเอียดจะเห็นได้ว่า การนำค่าประมาณของค่าสูญหายของตัวแปรตาม (Y) ที่ได้จากวิธีประมาณค่าสูญหายด้วยการถดถอย โดยมีตัวแปร X เป็นตัวแปรช่วย ไปใช้ในการวิเคราะห์การถดถอยที่มีตัวแปร Y เป็นตัวแปรตามและตัวแปร X เป็นตัวแปรอิสระ ทั้งนี้ข้อมูลของตัวแปร Y ส่วนหนึ่งเป็นข้อมูลที่ได้จากการสังเกตและอีกส่วนหนึ่งได้มาจากการประมาณค่าสูญหายดังกล่าว และตัวแปรอิสระ X คือ ตัวแปรช่วยที่ใช้ในการประมาณค่าสูญหายของตัวแปรตาม Y ซึ่งน่าจะมีผลทำให้ความแปรปรวนตัวอย่าง (Sample Variance) มีค่าน้อยกว่าที่ควรจะเป็น เกิดปัญหาการประมาณค่าต่ำ (Underestimation) ในทำนองเดียวกับวิธีประมาณค่าสูญหายที่ใช้ในรูปแบบค่าคงที่ (วิธีประมาณค่าสูญหายด้วยค่าเฉลี่ย วิธี RQ1 และวิธี RQ3) วิธี RRQ1 และวิธี RRQ3 แม้จะนำค่าของตัวแปร X มาใช้ทั้งในส่วนของตัวแปรช่วยและตัวแปรตาม แต่ลักษณะการใช้ในตัวแปรช่วย แต่ยังมีการปรับค่าด้วยอัตราส่วน ดังนั้น หากงานวิจัยที่ต้องการวิเคราะห์การถดถอยที่จำเป็นต้องประมาณค่าสูญหายของตัวแปรตามด้วย ก็อาจพิจารณาเลือกวิธีการประมาณค่าสูญหายที่ผู้วิจัยได้นำเสนอ 2 วิธีนี้ (RRQ1 และ RRQ3) ไปใช้ให้เป็นประโยชน์ได้

7.2 ด้านการวิจัยครั้งต่อไป

การนำเสนอวิธีประมาณค่าสูญหายในครั้งนี้ อาจยังขาดรายละเอียดในเชิงทฤษฎี และรายละเอียดในหลายเรื่อง เช่น การกระจายของข้อมูลทั้งก่อนและหลังการประมาณค่าสูญหาย รวมทั้งเกณฑ์การวัดประสิทธิภาพ ดังนั้น ในการศึกษาครั้งต่อไป อาจนำผลการศึกษาครั้งนี้เป็น ‘Baseline’ สำหรับการศึกษาในรายละเอียดและการศึกษาที่ลึกซึ้งมากขึ้น เช่น การกำหนดค่าความแปรปรวนของความคลาดเคลื่อนและค่าความแปรปรวนของตัวแปรอิสระให้ต่างออกไปจากงานวิจัยครั้งนี้ เพราะความแปรปรวนของตัวแปรอิสระหรือตัวแปร X จะมีผลต่อค่าควอไทล์ที่ 1 และค่าควอไทล์ที่ 3 ซึ่งน่าจะมีผลต่อตัวประมาณที่ผู้วิจัยได้เสนอแนะ รวมทั้งควรศึกษาเกณฑ์ที่ใช้ในการเปรียบเทียบเพิ่มเติมจากเกณฑ์ที่ใช้ในงานวิจัยครั้งนี้ด้วย

เอกสารอ้างอิง

จริยา แสงสุวรรณ. (2551). *การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ.* วิทยานิพนธ์มหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์.

ชุติมา ชัยมุสิก. (2533). *การวิเคราะห์การถดถอยเชิงซ้อนเมื่อข้อมูลของตัวแปรอิสระสูญหาย.* วิทยานิพนธ์มหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

ประชุม สุวัตถี. (2552). *การสำรวจด้วยตัวอย่าง: การชักตัวอย่างและการวิเคราะห์.* กรุงเทพฯ: โครงการส่งเสริมและพัฒนาเอกสารวิชาการ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

ประชุม สุวัตถี. (2554). *ทฤษฎีการชักตัวอย่าง.* กรุงเทพฯ: โครงการส่งเสริมและพัฒนาเอกสารวิชาการ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

วารุณี ตรีบำรุงศักดิ์. (2538). *การพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุเมื่อตัวแปรตามมีค่าสูญหาย.* วิทยานิพนธ์มหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

Abu-Dayyeh W.A., Ahmed M.S., Ahmed R.A. and Muttlak H.A. (2003). Some estimators of a finite population mean using auxiliary information. *Applied Mathematics and Computation.* 139, 287-298.

Al-omari A.I., Jemain A.A. and Ibrahim K. (2009). New ratio estimators of the mean using simple random sampling and ranked set sampling methods. *Revista investigacion operacional.* 30, 97-108.

Christine Bono and others. (2007). Missing data on the Center for Epidemiologic Studies Depression Scale: A comparison of 4 imputation techniques. *Research in Social and Administrative Pharmacy.* 3, 1-27.

Cochran, W.G. (1977). Sampling Techniques, 3rd edition, New York: Wiley.

Kadilar, C. and Cingi, H. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Compution.* 151(3), 893-902.

Laaksonen, S. (2000). Regression-Based Nearest Neighbor Hot Decking. *Computation Statistics.* 15, 65-71.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Nuanpan Nangsue. (2009). Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing. *World Academy of Science.* 29.

Ray, S.K. and Singh, R.K. (1981). Difference-cum-ratio type estimators. *Journal of Indian Statistical Association.* 19(24), 147-151.

Singh, H.P. and Kakran, M.S. (1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. Unpublished.

Singh, H.P. and Tailor, R. (2003). Use of Known Correlation Coefficient in Estimating the Finite Population Mean. *Statistics in Transition.* 6, 555-560.

Sisodia, B.V.S. and Dwivedi, V.K. (1981). A modifiedratio estimator using coefficient of variation of auxiliary variable. *Journal of Indian Society Agricultural Statistics.* 33, 13-18.

Translated Thai References

Chariya Saengsuwan. (2008). *A Comparative Study of Missing Data Estimation Methods in Multiple Regression Analysis.* Master Science (Statistics). Bangkok: Kasetsart University. (In Thai)

Chutima Chaimusik. (1990). *Missing Data in Multiple Regression Analysis.* Master Science (Statistics). Bangkok: Chulalongkorn University. (In Thai)

Prachoom Suwattee. (2009). *Sample Surveys: Sampling Design and Analysis.* Bangkok: National Institute of Development Administration. (In Thai)

Prachoom Suwattee. (2011). *Sampling Theory.* Bangkok: National Institute of Development Administration. (In Thai)

Warunee Treebumrungsak. (1995). *Forecasting in Multiple Linear Regression with Missing Observations in the Dependent Variable.* Master Science (Statistics). Bangkok: Chulalongkorn University. (In Thai)

ภาคผนวก

ตารางที่ ก.1: ค่า MAPE เมื่อค่าความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5

ขนาดตัวอย่าง	ค่าสูญหาย (%)	วิธีประมาณค่าสูญหาย					
		Mean	Regression	RQ1	RQ3	RRQ1	RRQ3
20	10	18.20081	9.48845	18.01594	18.05371	9.48327	9.48349
	15	27.94389	13.57058	27.77708	27.81909	13.59827	13.59368
	20	29.81033	13.80985	29.23210	29.34063	13.73854	13.74771
40	10	16.08161	5.48670	15.88162	15.92039	5.50593	5.50167
	15	23.52623	7.09308	23.36096	23.39477	7.08126	7.08334
	20	26.08396	9.15485	25.84606	25.88972	9.15017	9.14893
60	10	15.97179	4.68592	15.86682	15.88702	4.68147	4.68170
	15	24.24839	7.63947	24.09851	24.12734	7.62888	7.63091
	20	31.57739	9.16628	31.25051	31.30321	9.28025	9.26290
100	10	11.18051	3.57242	11.11708	11.12899	3.57112	3.57111
	15	19.27245	4.32358	19.22859	19.23721	4.33188	4.32969
	20	28.40513	5.45506	28.76869	28.70985	5.42509	5.42877

ตารางที่ ก.2: ค่า MAPE เมื่อค่าความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1

ขนาดตัวอย่าง	ค่าสูญหาย (%)	วิธีประมาณค่าสูญหาย					
		Mean	Regression	RQ1	RQ3	RRQ1	RRQ3
20	10	38.66492	24.94750	38.53323	38.56138	24.95297	24.95120
	15	53.27186	30.79129	53.33831	53.33449	30.88059	30.86402
	20	56.66253	34.64248	56.21369	56.29525	34.54218	34.55802
40	10	28.82947	14.39443	28.37828	28.45836	14.26844	14.28796
	15	45.00489	19.25736	44.41591	44.51969	19.16421	19.17998
	20	37.98397	19.21228	37.54952	37.62793	19.18864	19.19317
60	10	13.71474	6.56442	13.63909	13.65360	6.55580	6.55712
	15	20.65494	8.51071	20.54488	20.56612	8.48860	8.49273
	20	28.12173	10.22065	27.99477	28.02053	10.21913	10.21938
100	10	12.49089	4.61417	12.43038	12.44167	4.61400	4.61374
	15	19.69526	6.48602	19.61662	19.63144	6.48560	6.48511
	20	26.58320	8.61403	26.87553	26.82888	8.64172	8.63584

ตารางที่ ก.3: ค่า MAPE เมื่อค่าความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1.5

ขนาดตัวอย่าง	ค่าสูญหาย (%)	วิธีประมาณค่าสูญหาย					
		Mean	Regression	RQ1	RQ3	RRQ1	RRQ3
20	10	38.00024	23.35377	37.57422	37.67354	23.44908	23.43346
	15	59.79342	32.36084	59.02352	59.20623	32.45460	32.43445
	20	66.69154	30.41961	64.66140	65.09981	30.21133	30.25211
40	10	24.99724	12.46728	24.79795	24.83780	12.45661	12.45794
	15	46.34170	22.23529	45.97337	46.05780	22.17004	22.18442
	20	55.63172	31.82949	55.21172	55.29339	31.82934	31.82649
60	10	15.52098	9.15747	15.43220	15.44968	9.15901	9.15852
	15	23.43602	10.77639	23.29990	23.32604	10.76585	10.76751
	20	32.29011	12.79971	32.13112	32.16306	12.78845	12.79009
100	10	15.40272	6.78762	15.33313	15.34601	6.78386	6.78428
	15	23.30502	9.60801	23.22493	23.23976	9.60276	9.60334
	20	32.37109	10.61091	32.68533	32.63650	10.64722	10.63987

ตารางที่ ก.4: ค่า MAPE เมื่อค่าความแปรปรวนของความคลาดเคลื่อนเท่ากับ 2

ขนาดตัวอย่าง	ค่าสูญหาย (%)	วิธีประมาณค่าสูญหาย					
		Mean	Regression	RQ1	RQ3	RRQ1	RRQ3
20	10	37.95029	24.74611	37.65057	37.71311	24.70551	24.71334
	15	49.94820	31.07876	49.75309	49.80477	31.09947	31.09946
	20	57.33338	31.71277	56.44624	56.60958	31.54778	31.57538
40	10	18.00197	12.12684	17.87764	17.90344	12.13262	12.12983
	15	28.21087	19.60226	28.10052	28.12368	19.58914	19.59143
	20	34.05896	19.80577	33.70757	33.77639	19.79437	19.79188
60	10	20.83830	17.88690	20.92845	20.91165	17.96649	17.94871
	15	32.39106	18.78922	32.21883	32.25318	18.79892	18.79723
	20	42.81256	24.09577	42.56855	42.61932	24.05321	24.06175
100	10	18.59255	8.44431	18.48048	18.50137	8.42817	8.43116
	15	25.21082	12.94307	25.11133	25.12937	12.95515	12.95235
	20	41.46723	14.08079	41.87239	41.80836	14.15342	14.14099