การเปรียบเทียบวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

ศิริวัฒนา สีดี

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2564

การเปรียบเทียบวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

ศิริวัฒนา สีดี

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2564

วิทยานิพนธ์ เรื่อง "การเปรียบเทียบวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ"

ของ ศิริวัฒนา สีดี

ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติ

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์

(ศาสตราจารย์ ดร.ยุพาภรณ์ อารีพงษ์)

..... ประธานที่ปรึกษาวิทยานิพนธ์

(รองศาสตราจารย์ ดร.เกตุจันทร์ จำปาไชยศรี)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(ดร.สวพร หิญชีระนันทน์)

อนุมัติ

.....

(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)

คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

ผู้วิจัย ศิริวัฒนา สีดี

ประธานที่ปรึกษา รองศาสตราจารย์ ดร.เกตุจันทร์ จำปาไชยศรี

ประเภทสารนิพนธ์ วิทยานิพนธ์ วท.ม. สถิติ, มหาวิทยาลัยนเรศวร, 2564

คำสำคัญ การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ, ข้อมูลสูญหาย, วิธีแบบเบส์, วิธีเบส์เซียนบูตสแตรป

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าตัวแปรตามที่สูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) และวิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์ (Bay-in) วิธีแบบเบส์ที่ให้สารสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI) ทำการจำลองข้อมูลโดยใช้โปรแกรม R กระทำซ้ำ 10,000 ครั้ง โดยใช้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ กำหนดขนาดตัวอย่างที่ใช้ในการศึกษาเป็น 50, 100 และ 200 ระดับความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเป็น ต่ำ ปานกลาง และสูง เปอร์เซ็นต์การสูญหายของข้อมูลในตัวแปรตามเป็น 5, 10 และ 20 ความแปรปรวนของความคลาดเคลื่อนเป็น 0.5, 1, 2, 5 และ 10 ผลการวิจัยพบว่า ในทุกระดับขนาดตัวอย่างและเปอร์เซ็นต์การสูญหายของข้อมูล เมื่อความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามอยู่ในระดับต่ำและปานกลาง ความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5 พบว่า วิธี Bay-in มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1 และ 2 วิธี Bay-in และวิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ แต่เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5 และ 10 พบว่า วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ กรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง พบว่าวิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกค่าความแปรปรวนของความคลาดเคลื่อน

Title A COMPARISON OF MISSING RESPONSE ESTIMATION METHODS IN MULTIPLE LINEAR REGRESSION

Author SIRIWATTANA SEEDEE

Advisor Associate Professor Dr. Katechan Jampachaisri

Academic Paper M.S. Thesis in Statistics - (Type A 2), Naresuan University, 2021

Keywords Multiple Linear Regression Analysis Regression Imputation method Multiple Imputation method Expectation Maximization Algorithm method Bayesian Method Bayesian Bootstrap

ABSTRACT

The objective of this research is to compare the efficiency of 6 estimation methods for missing response in multiple linear regression: Regression Imputation method (RI), Multiple Imputation method (MI), Expectation Maximization Algorithm method (EM), Bayes' method with informative prior (Bay-in), Bayes' method with non-informative priors (Bay-non) and Bayes bootstrap regression imputation method (BBRI). Data are simulated and repeated 10,000 times using R program. The average mean squares error (AMSE) is used as criteria for comparison. The sample sizes were set to 50, 100 and 200, with low, moderate and high levels of associations between independent variables and response, the percentage of missing response as 5, 10 and 20, and error variances as 0.5, 1, 2, 5 and 10. The results indicated that, for all sample sizes all percentages of missing response with low and moderate associations and error variance equaling 0.5, Bay-in was the most efficient in most cases. When error variances equal to 1 and 2, Bay-in and BBRI were mostly efficient while BBRI was the most efficient as error variances equal to 5 and 10. With high association, BBRI was mostly efficient in all values of error variances.

ประกาศคุณูปการ

การวิจัยเรื่องการเปรียบเทียบวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ สำเร็จลุล่วงไปด้วยดีโดยได้รับความอนุเคราะห์และความช่วยเหลือจากอาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร.เกตุจันทร์จำปาไชยศรีที่ได้เสียสละเวลา อีกทั้งให้คำปรึกษา แนะนำ ตรวจสอบและแก้ไขข้อบกพร่องต่าง ๆ เพื่อให้งานวิจัยฉบับนี้มีความสมบูรณ์ ซึ่งผู้วิจัยขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ศ.ดร.ยุพาภรณ์อารีพงษ์ และ ดร.สวพร หิญชีระนันทน์ กรรมการสอบวิทยานิพนธ์ที่ให้คำแนะนำและข้อเสนอแนะต่าง ๆ ในงานวิจัย เพื่อนำมาปรับปรุงและแก้ไขให้งานวิจัยฉบับนี้สมบูรณ์ยิ่งขึ้น

ท้ายที่สุดนี้ ผู้วิจัยหวังเป็นอย่างยิ่งว่าผลของการศึกษาจะเป็นประโยชน์ไม่มากก็น้อยสำหรับหน่วยงานที่เกี่ยวข้อง ตลอดจนผู้สนใจ หากมีข้อผิดพลาดประการใด ผู้วิจัยขออภัยเป็นอย่างสูง

ศิริวัฒนา สีดี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของการศึกษา

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple linear regression analysis) เป็นการวิเคราะห์ข้อมูลโดยใช้ตัวแปรอิสระ (Independent variable) ที่มีมากกว่าหนึ่งตัวขึ้นไปมาใช้ในการอธิบายตัวแปรตาม (Dependent variable) ซึ่งตัวแปรอิสระและตัวแปรตามจะมีความสัมพันธ์กันในลักษณะใดลักษณะหนึ่ง ปัญหาหนึ่งที่อาจเกิดขึ้นในการวิเคราะห์การถดถอยคือ การสูญหายของข้อมูล (Missing data) ในตัวแปรตาม ตัวอย่างเช่น การเก็บข้อมูลจากการสำรวจ อาจเกิดปัญหาข้อมูลสูญหายเนื่องจากหน่วยตัวอย่างบางหน่วยไม่ให้คำตอบในบางประเด็นคำถาม หน่วยตัวอย่างนี้ทราบข้อมูลที่เป็นค่าของตัวแปรอิสระที่ทำการศึกษา แต่ไม่ทราบข้อมูลที่เป็นค่าของตัวแปรตาม ซึ่งคือคำตอบของประเด็นคำถามนั้น หรือตัวอย่างข้อมูลจากการทดลองที่เกิดการสูญหาย เนื่องจากหน่วยทดลองเป็นสิ่งมีชีวิตแล้วเกิดการตายในระหว่างการทดลอง หน่วยตัวอย่างนี้ทราบข้อมูลที่เป็นค่าของตัวแปรอิสระแต่ไม่ทราบข้อมูลที่เป็นค่าของตัวแปรตามเนื่องจากเกิดการตาย ในกรณีที่ไม่สามารถเก็บข้อมูลเพิ่มได้อันเนื่องจากมีข้อจำกัดในเรื่องของเวลา หรือมีข้อจำกัดจากปัจจัยอื่นๆ ซึ่งถือเป็นปัญหาที่มีความสำคัญในการวิเคราะห์ข้อมูล หากนำข้อมูลที่ไม่สมบูรณ์ไปวิเคราะห์ อาจทำให้สูญเสียอำนาจการทดสอบ ทำให้ได้ผลลัพธ์ที่คลาดเคลื่อนไปจากความเป็นจริง ส่งผลกระทบต่อการนำไปใช้ในการวางแผนตัดสินใจในงานต่างๆ อีกทั้งยังทำให้ประสิทธิภาพในการวิเคราะห์ข้อมูลลดลง โดยผลลัพธ์ที่ได้อาจเกิดความเอนเอียง (Bias) (เรืองลักษณ์ หลำใจซื่อ, อำไพ ทองธีรภาพ และจุฑาภรณ์ สินสมบูรณ์ทอง, 2560)

การจัดการกับปัญหาข้อมูลสูญหายนั้นสามารถทำได้หลายวิธี และวิธีที่ง่ายที่สุดคือ การตัดข้อมูลที่สูญหายทิ้ง และนำข้อมูลที่สมบูรณ์เท่านั้นมาวิเคราะห์ หากข้อมูลสมบูรณ์ที่มีอยู่มีจำนวนมากพอที่จะนำมาวิเคราะห์ก็สามารถตัดข้อมูลที่สูญหายทิ้งไปได้ (Ignoring and discarding data) (นรุตม์ บุตรพลอย, 2553) แต่บางครั้งข้อมูลสมบูรณ์ที่เหลืออยู่มีจำนวนไม่มาก จึงไม่เพียงพอต่อการวิเคราะห์ (สุปรียา สระโสม และ ธิดาเดียว มยุรีสวรรค์, 2547) ทำให้มีผู้คิดค้นศึกษา และพัฒนาวิธีการจัดการข้อมูลสูญหาย เพื่อเพิ่มประสิทธิภาพในการประมาณค่าสูญหายที่ดีกว่าวิธีการเดิม การศึกษางานวิจัยที่เกี่ยวข้องกับการประมาณค่าสูญหายของตัวแปรตาม มีดังนี้ Brandel (2004) ได้

ศึกษาการสูญหายของข้อมูลในการทดลองทางการแพทย์ ซึ่งมีสาเหตุหลายประการ เช่น ผู้ป่วยบางส่วนที่เข้าร่วมในการทดลอง ปฏิเสธที่จะทำการทดลองต่อ หรือผู้ป่วยไม่ปฏิบัติตามข้อกำหนด ทำให้เกิดความผิดพลาดในการวิเคราะห์ แสดงให้เห็นว่า การทดลองไม่เป็นไปตามหลักการศึกษา ซึ่งปัญหาเหล่านี้ไม่สามารถเพิกเฉยได้ โดยงานวิจัยนี้ได้ทำการศึกษาการประมาณค่าสูญหายด้วยวิธีการแบบเบส์ วิธีค่าคาดหวังสูงสุด (EM) และเปรียบเทียบกับวิธีโดยทั่วไปอีก 4 วิธี ได้แก่ วิธี Last Observation Carried Forward (LOCF) วิธีสมการถดถอย (Regression Imputation) วิธี Best or worst case imputation และวิธีค่าเฉลี่ย (Mean) ผลการวิจัยพบว่า เมื่อเปอร์เซ็นต์การสูญหายเป็น 10% ค่าความแปรปรวนลดลง วิธีการแบบเบส์ให้ค่าประมาณได้ดีกว่าวิธีอื่น และเมื่อเปอร์เซ็นต์การสูญหายเป็น 30% พบว่าวิธี EM เริ่มประมาณค่าได้ดีขึ้น เมื่อเปอร์เซ็นต์การสูญหายเป็น 50% วิธีการแบบเบส์ประมาณค่าได้ดีกว่าทุกวิธี นอกจากนี้ยังสรุปได้ว่าวิธี EM ประมาณค่าได้ดีกว่าวิธีโดยทั่วไปอีก 4 วิธีที่เหลืออย่างมีนัยสำคัญ สำหรับเปอร์เซ็นต์การสูญหายเป็น 80% พบว่าวิธีการแบบเบส์และวิธี EM ให้ค่าประมาณที่คล้ายกันมาก ทั้งนี้ สามารถสรุปได้ว่าวิธีการแบบเบส์ให้ค่าประมาณของข้อมูลได้ดีกว่าวิธีอื่นในทุกเปอร์เซ็นต์สูญหาย และทุกๆค่าพารามิเตอร์ของข้อมูล เช่น ค่าเฉลี่ย ความแปรปรวน จำนวนข้อมูลผู้ป่วยที่สังเกตได้ สำหรับวิธี EM สามารถสรุปได้เช่นเดียวกับวิธีการแบบเบส์ แต่มีข้อจำกัดว่าเปอร์เซ็นต์สูญหายของข้อมูลต้องมากกว่า 30% ขึ้นไป จึงจะประมาณค่าได้ดี จริยา แสงสุวรรณ, (2551) ได้เปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในตัวแปรตาม สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยศึกษา 4 วิธีได้แก่ วิธี Loss Imputation (Loss) วิธี Mean Imputation (Mean) วิธี Regression Imputation (RI) และวิธี Multiple Imputation (MI) ผลการวิจัยพบว่า เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธี RI และวิธี MI ให้ค่าประมาณของ Root Mean Square Error (RMSE) ลดลง และวิธีการประมาณค่าสูญหายทั้ง 4 วิธี ให้ค่าประมาณของ RMSE แตกต่างกัน วิธี RI และวิธี MI ให้ค่าประมาณของ RMSE ใกล้เคียงกัน แต่เนื่องจากวิธี RI เป็นวิธีที่ง่ายและไม่ซับซ้อน ดังนั้น จึงเป็นวิธีที่เหมาะสมในการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ Erlangung (2009) ศึกษาการวิเคราะห์ข้อมูลการสำรวจที่มีค่าขาดหายไป โดยได้ศึกษาวิธี Bayesian Bootstrap Predictive Mean Matching (BBPMM) เปรียบเทียบกับวิธี Posterior Predictive Mean Matching (PPMM), Rounded Predictive Mean Matching (RPMM) และวิธี Rounding to the nearest observed value (ROV) ผลการวิจัยพบว่า โดยรวมแล้ว วิธี PPMM และ BBPMM มีค่าความเอนเอียงเฉลี่ยน้อยที่สุด แต่วิธี ROV เป็นวิธีที่มีค่าความเอน

เอียงเฉลี่ยมากที่สุด เมื่อพิจารณาค่าความครอบคลุมเฉลี่ย (Average coverage) โดยกำหนดช่วงความเชื่อมั่น 95% พบว่า วิธี BBPMM มีค่าเข้าใกล้ 95% มากที่สุดเป็นส่วนใหญ่ ต่อมา ธรรมรัตน์ กลีบเมฆ และนพคุณ ทองมวล, (2563) ศึกษาการประมาณค่าข้อมูลสูญหายเมื่อตัวแปรตาม Y มีความสัมพันธ์กับตัวแปรอิสระ X และตัวแปร X และ Y มีการแจกแจงปรกติ โดยนำเสนอวิธีประมาณค่าข้อมูลสูญหาย ด้วยวิธีการถดถอยแบบเบส์-บูตสแตรป และทำการเปรียบเทียบกับวิธีประมาณค่าสูญหายด้วยวิธีถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด โดยใช้เกณฑ์ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเพื่อวัดความแม่นยำ ผลการศึกษาพบว่า วิธีการถดถอยแบบเบส์-บูตสแตรปและวิธีการถดถอยมีความแม่นยำมากกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุดในทุกกรณี และมีบางกรณีที่ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอยแบบเบส์-บูตสแตรปมีค่าต่ำสุด จากงานวิจัยที่กล่าวมาข้างต้น ผู้วิจัยมีความสนใจที่จะศึกษาวิธีการถดถอยแบบเบส์บูตสแตรป ซึ่งถูกเสนอโดย ธรรมรัตน์ กลีบเมฆ และนพคุณ ทองมวล, (2563) ที่ศึกษาและพัฒนาวิธีการประมาณค่าข้อมูลสูญหายในตัวแปรตาม เมื่อตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์กัน โดยได้นำวิธีเบส์บูตสแตรป (Rubin,1981) มาประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเพื่อใช้ประมาณค่าตัวแปรตามที่สูญหาย พบว่า วิธีการถดถอยแบบเบส์บูตสแตรปมีประสิทธิภาพมากที่สุดเป็นส่วนใหญ่ในทุกกรณีศึกษา โดยนำมาเปรียบเทียบกับวิธีการแบบเบส์ ในงานวิจัยของ Brandel (2004) ผลการวิจัยพบว่า วิธีการแบบเบส์ให้ค่าประมาณของข้อมูลได้ดีกว่าวิธีอื่นในทุกเปอร์เซ็นต์สูญหาย และทุกๆค่าพารามิเตอร์ของข้อมูล

งานวิจัยครั้งนี้ ผู้วิจัยได้นำวิธีการถดถอยแบบเบส์บูตสแตรป (Bayes bootstrap regression Imputation : BBRI) วิธีแบบเบส์ที่ให้สาระสนเทศที่เป็นประโยชน์ (Bay-in) และได้มีการเพิ่มวิธีแบบเบส์ที่ให้สารสนเทศน้อยมาก (Bay-non) เข้ามาในงานวิจัยนี้ด้วย โดยนำมาเปรียบเทียบกับอีก 3 วิธี คือ วิธีสมการถดถอย (Regression Imputation : RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (Multiple Imputation : MI) และวิธีค่าคาดหวังสูงสุด (Expectation Maximization Algorithm : EM) โดยศึกษาในกรณีที่ตัวแปรตามมีการสูญหายแบบสุ่ม (MAR) ซึ่งเป็นการสูญหายที่มีความน่าจะเป็นขึ้นอยู่กับตัวแปรตัวอื่นที่ทราบค่า และใช้ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error : AMSE) เป็นเกณฑ์ในการเปรียบเทียบ

1.2 วัตถุประสงค์การวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) วิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์ (Bay-in) วิธีแบบเบส์ที่ให้สารสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI)

1.3 ขอบเขตของการวิจัย

1.3.1 กำหนดตัวแปรอิสร 2 ตัว มีการแจกแจงปรกติ นั่นคือ $X_1 \sim N(0,1)$ และ $X_2 \sim N(0,10)$ โดยที่ตัวแปรอิสระไม่มีการสูญหายและมีความสัมพันธ์กันต่ำ

1.3.2 กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 0.5, 1, 2, 5, 10 และกำหนดขนาดตัวอย่างที่ศึกษาเป็น 50, 100 และ 200

1.3.3 สร้างตัวแปรตามให้มีความสัมพันธ์กับตัวแปรอิสระ โดยกำหนดระดับความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม 3 ระดับ คือ (ธรรมรัตน์ กลีบเมฆ และนพคุณ ทองมวล, 2563)

1. ระดับต่ำ $0 \le \rho < 0.4$
2. ระดับปานกลาง $0.4 \le \rho < 0.7$
3. ระดับสูง $0.7 \le \rho < 1.0$

1.3.4 กำหนดให้ตัวแปรตามให้มีการสูญหายแบบสุ่ม (MAR) มีเปอร์เซ็นต์การสูญหายอยู่ที่ 5, 10 และ 20 ของขนาดตัวอย่างที่ศึกษา โดยกำหนดให้การสูญหายขึ้นอยู่กับตัวแปร X_2

1.3.5 กำหนดการแจกแจงก่อนที่ให้สารสนเทศน้อยมาก $f\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$

1.3.6 กำหนดการแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ $\boldsymbol{\beta} \sim N\left(\boldsymbol{A}, \boldsymbol{\Sigma}\right)$ และ $\sigma^2 \sim Inv-gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$ เมื่อ $\boldsymbol{A}, \boldsymbol{\Sigma}, \nu_0, \sigma_0^2$ แทน ไฮเปอร์พารามิเตอร์ที่ทราบค่า

1.3.7 กำหนดจำนวนรอบบูตสแตรป เท่ากับ 1,000 รอบ

1.3.8 กำหนดจำนวนรอบของวิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) จำนวน 5 รอบ

1.3.9 การศึกษานี้ได้ทำการจำลองค่าตัวแปรสุ่มตามการแจกแจงของประชากรที่กำหนดและทำซ้ำ 10,000 รอบ ในแต่ละสถานการณ์ โดยใช้โปรแกรม R studio ในการวิเคราะห์ข้อมูล

1.4 คำสำคัญในงานวิจัย

1.4.1 การแจกแจงก่อน (Prior Distribution) หมายถึง การสำรวจข้อมูลล่วงหน้า เพื่อนำลักษณะของข้อมูล เช่น ค่าเฉลี่ย ความแปรปรวน ไปปรับปรุงตัวประมาณค่าพารามิเตอร์สำหรับการประมาณค่าพารามิเตอร์แบบเบส์

1.4.2 การแจกแจงภายหลัง (Posterior Distribution) หมายถึง ข้อมูลในอนาคตที่เป็นผลจากการทราบข้อมูลปัจจุบันและอดีต

1.4.3 การสูญหายแบบสุ่ม (Missing at random หรือ MAR) เป็นลักษณะการสูญหายของข้อมูลที่ความน่าจะเป็นของค่าสังเกตที่สูญหายขึ้นอยู่กับค่าของตัวแปรอื่นๆที่ทราบค่า แต่จะไม่ขึ้นอยู่กับค่าของข้อมูลที่สูญหาย (Little and Rubin, 2002)

1.5 เกณฑ์ในการตัดสินใจ

การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จะพิจารณาจากค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) โดยวิธีที่ให้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่าจะพิจารณาว่าเป็นวิธีที่มีประสิทธิภาพมากกว่า

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 เป็นแนวทางในการเลือกวิธีการประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

1.6.2 เป็นแนวทางในการศึกษาวิธีประมาณค่าสูญหายวิธีอื่น

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) วิธีเบส์ที่ให้สาระสนเทศที่เป็นประโยชน์ (Bay-in) วิธีเบส์ที่ให้สาระสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI) มีทฤษฎีและงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดในการนำเสนอ ดังนี้

2.1 รูปแบบของข้อมูลสูญหาย

2.2 ประเภทของข้อมูลสูญหาย

2.3 ข้อตกลงเบื้องต้นในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

2.4 ตัวแบบถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression)

2.5 ทฤษฎีที่เกี่ยวข้อง

2.6 วิธีการประมาณค่าข้อมูลสูญหาย

2.7 เกณฑ์การตัดสินใจ

2.8 งานวิจัยที่เกี่ยวข้อง

2.1 รูปแบบของข้อมูลสูญหาย

รูปแบบของข้อมูลสูญหายมีหลากหลายรูปแบบ มีทั้งข้อมูลสูญหายเพียงตัวแปรเดียวและสูญหายหลายตัวแปร สามารถแบ่งประเภทได้ดังนี้ (พิมพ์ชนก เชาวณาพรรณ์, 2559)

1. รูปแบบที่มีข้อมูลสูญหายตัวแปรเดียว (Univariate Pattern) คือ มีข้อมูลสูญหายเพียง 1 ตัวแปรมักพบได้ในข้อมูลจากการวางแผนการทดลอง

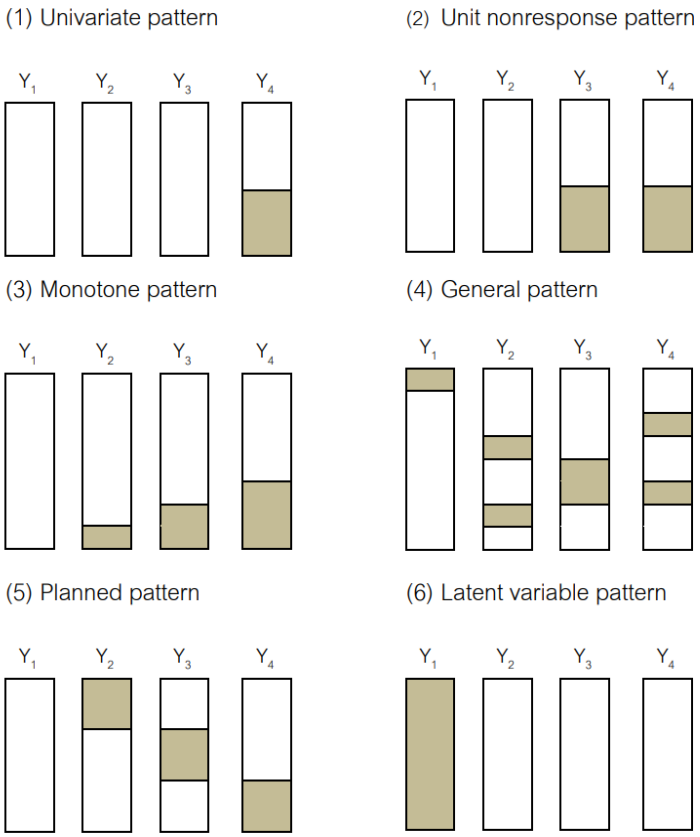
2. รูปแบบที่มีข้อมูลสูญหายมากกว่าตัวแปร (Unit Nonresponse Pattern) คือ ในข้อมูลช่วงเดียวกัน มีค่าของตัวแปรที่สูญหายมากกว่า 1 ตัว

3. รูปแบบที่มีข้อมูลสูญหายไปในทิศทางเดียวกัน (Monotone Pattern) คือ มีการสูญหายของข้อมูลเพิ่มขึ้นตามลำดับ นั่นคือ ตัวแปร $\boldsymbol{Y}_1$ จะมีข้อมูลมากกว่าตัวแปร $\boldsymbol{Y}_2$ และตัวแปร $\boldsymbol{Y}_2$ จะมีข้อมูลมากกว่าตัวแปร $\boldsymbol{Y}_3$ มักพบได้ในข้อมูลที่เก็บมาในระยะยาว

4. รูปแบบที่มีข้อมูลสูญหายที่เกิดขึ้นได้ทั่วไป (General Pattern) คือ มีข้อมูลสูญหายอย่างไม่มีระบบ กระจายตัวในบางช่วงของตัวแปรหรือบางช่วงของข้อมูลในชุดเดียวกัน

5. รูปแบบที่มีข้อมูลสูญหายเป็นแบบแผน (Planned Pattern) คือ ไม่มีข้อมูลสูญหายในชุดเดียวกัน

6. รูปแบบที่มีข้อมูลสูญหายเป็นตัวแปรหนึ่ง (Latent Variable Pattern) คือ มีตัวแปรใดตัวแปรหนึ่งเป็นข้อมูลที่สูญหายทั้งหมด หรือไม่มีค่าสังเกตในตัวแปรนั้นเลย



ภาพที่ 2.1 รูปแบบการสูญหายของข้อมูล

2.2 ประเภทของข้อมูลสูญหาย

การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญ หากทราบลักษณะของข้อมูลสูญหายจะช่วยในการพิจารณาหาแนวทางเพื่อจัดการกับปัญหาความไม่สมบูรณ์หรือการสูญหายของข้อมูลได้อย่างเหมาะสม ซึ่งโดยทั่วไปสามารถจำแนกข้อมูลสูญหายออกเป็น 3 ประเภท ดังนี้

1. การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random : MCAR) เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด โดยความน่าจะเป็นของการเกิด

ข้อมูลสูญหายไม่ขึ้นอยู่กับค่าของตัวแปรอื่นๆ ไม่ว่าจะเป็นตัวแปรที่ทราบค่าหรือไม่ทราบค่าก็ตาม สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้ได้ โดยแบ่งกลุ่มค่าสังเกตออกเป็นกลุ่มข้อมูลที่สมบูรณ์และข้อมูลที่สูญหาย เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างกลุ่ม สาเหตุที่ทำให้ข้อมูลเกิดการสูญหายประเภทนี้มีอยู่หลากหลายเหตุผล ซึ่งอาจเกิดจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้อง

2. การสูญหายแบบสุ่ม (Missing at Random : MAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มกับบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือค่าข้อมูลสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในฐานข้อมูล ซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย เช่น ในการเก็บรวบรวมข้อมูลหากพบว่าเฉพาะกลุ่มผู้มีรายได้มากที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการใช้จ่าย ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการใช้จ่ายมีค่าสูญหายแบบ MAR เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นกับเฉพาะบางกลุ่มรายได้

3. การสูญหายแบบไม่สุ่ม (Not Missing at Random : NMAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกันรวมถึงตัวแปรตัวอื่นด้วย เช่น ข้อมูลสูญหายของระดับรายได้ขึ้นอยู่ระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นนี้จัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใด ๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้น แต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูล

2.3 ข้อตกลงเบื้องต้นในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์เชิงเส้นระหว่างตัวแปรตาม และตัวแปรอิสระตั้งแต่ 2 ตัวแปรขึ้นไป มีข้อตกลงเบื้องต้น ดังนี้

1. ความคลาดเคลื่อน ε_i เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปรกติ

2. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์ นั่นคือ $E(\varepsilon_i)=0$

3. ความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ $Var(\varepsilon_i)=\sigma^2$

4. ε_i และ ε_j เป็นอิสระต่อกัน นั่นคือ $Cov(\varepsilon_i,\varepsilon_j)=0$ โดยที่ $i \neq j$

5. X_i และ X_j เป็นอิสระต่อกัน เมื่อ $i, j = 1, 2, ..., k$

2.4 ตัวแบบถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression)

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ประกอบด้วยตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป โดยตัวแบบของการถดถอยเชิงเส้นพหุคูณ เมื่อมีตัวแปรอิสระ k ตัว สามารถแสดงได้ดังนี้

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \qquad i = 1, \ldots, n \qquad (2.1)$$

โดยที่

y_i แทน ค่าสังเกตของตัวแปรตาม

$\beta_0, \beta_1, ..., \beta_k$ แทน สัมประสิทธิ์ถดถอยซึ่งเป็นพารามิเตอร์ที่ไม่ทราบค่า

$X_{i1}, X_{i2}, ..., X_{ik}$ แทน ค่าสังเกตของตัวแปรอิสระ

ε_i แทน ความคลาดเคลื่อนสุ่ม

k แทน จำนวนตัวแปรอิสระในสมการถดถอย

หรือเขียนในรูปเมทริกซ์ได้ดังนี้ $\boldsymbol{Y} = \boldsymbol{X}\underset{\sim}{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$ โดย

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \underset{\sim}{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2.5 ทฤษฎีที่เกี่ยวข้อง

2.5.1 วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares: OLS)

การประมาณค่าสัมประสิทธิ์ถดถอยด้วยกำลังสองน้อยที่สุด เป็นการหาค่าประมาณของพารามิเตอร์ที่ทำให้ผลบวกกำลังสองของผลต่างระหว่างค่าสังเกตกับค่าคาดหวังของตัวแปรมีค่าต่ำสุด

$$\boldsymbol{Y} = \boldsymbol{X}\underset{\sim}{\hat{\boldsymbol{\beta}}} + \boldsymbol{\varepsilon} \qquad (2.2)$$

เมื่อ $\boldsymbol{Y}$ แทน เมทริกซ์ของตัวแปรตามขนาด $(n \times 1)$

$\boldsymbol{X}$ แทน เมทริกซ์ของตัวแปรอิสระขนาด $(n \times (k+1))$

$\underset{\sim}{\boldsymbol{\beta}}$ แทน เมทริกซ์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $((k+1) \times 1)$

n แทน ขนาดตัวอย่าง

p แทน จำนวนตัวแปรอิสระ โดย $p = k + 1$

$\boldsymbol{\varepsilon}$ แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $(n \times 1)$

ได้ตัวประมาณสัมประสิทธิ์ถดถอย ดังนี้

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{2.3}$$

โดยที่ $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ และ $\sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$

ดังนั้นสมการถดถอยที่ใช้ในการพยากรณ์ คือ

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} \tag{2.4}$$

2.5.2 การแจกแจงอินเวอร์สแกมมา (Inverse gamma distribution)

ให้ X แทน ตัวแปรสุ่ม มีการแจกแจงอินเวอร์สแกมมา โดยมีฟังก์ชันหนาแน่นน่าจะเป็น (Probability density function) เขียนแทนด้วย $X \sim IG(\alpha, \beta)$ ดังนี้

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right) \tag{2.5}$$

เมื่อ α แทน พารามิเตอร์รูปร่าง (Shape parameter)

β แทน พารามิเตอร์ตำแหน่ง (Scale parameter)

โดยมี ค่าเฉลี่ย : $E(X) = \frac{\beta}{\alpha-1}$ เมื่อ $\alpha > 1$

ความแปรปรวน : $V(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ เมื่อ $\alpha > 2$

2.5.3 การแจกแจงแบบทีหลายตัวแปร (multivariate t distribution)

ให้ X มีการแจกแจงทีหลายตัวแปร เขียนแทนด้วย $X \sim mvt(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ โดยมีฟังก์ชันหนาแน่นน่าจะเป็น (Probability density function) ดังนี้

$$f(x;\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma\left[(v+p)/2\right]}{\Gamma(v/2)v^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left[1+\frac{1}{v}(x-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right]^{-(v+p)/2} \tag{2.6}$$

เมื่อ $\boldsymbol{\mu} = \left[\mu_1,\ldots,\mu_p\right]^T_{p\times 1}$ แทน เมทริกซ์ค่าเฉลี่ยขนาด $p\times 1$

$\boldsymbol{\Sigma}$ แทน เมทริกซ์ความแปรปรวนร่วมขนาด $p\times p$

v แทน ระดับองศาเสรี

2.5.4 การสุ่มตัวอย่างแบบกิบส์ (Gibbs sampling)

เป็นวิธีการสุ่มตัวอย่างอีกวิธีหนึ่งที่นิยมใช้กรณีที่ทราบการแจกแจงก่อน เมื่อกำหนดค่าของตัวแปรสุ่ม ซึ่งถูกเสนอโดย Geman (1984) มีขั้นตอนดังนี้ (อัชฌา อระวีพร, 2555)

1. กำหนดค่าเริ่มต้น $\boldsymbol{\theta}^0$ จากฟังก์ชันการแจกแจงก่อน

2. ทำซ้ำตามขั้นตอนต่อไปนี้

2.1 กำหนด $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t-1)}$

2.2 สร้างค่า $\boldsymbol{\theta}_j$ จาก $\boldsymbol{\theta}_j \sim f\left(\boldsymbol{\theta}_j \middle| \boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{j-1},\boldsymbol{\theta}_{j+1},...,\boldsymbol{\theta}_d,\boldsymbol{y}\right)$ เมื่อ $j = 1,2,...,d$

2.3 กำหนดให้ $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$ เพื่อใช้สร้างพารามิเตอร์ที่รอบ $t+1$ ตามกระบวนการนี้

$\boldsymbol{\theta}_1^{(t)}$ จาก $f\left(\boldsymbol{\theta}_1 \middle| \boldsymbol{\theta}_2^{(t-1)},\boldsymbol{\theta}_3^{(t-1)},...,\boldsymbol{\theta}_d^{(t-1)},\boldsymbol{y}\right)$

$\boldsymbol{\theta}_2^{(t)}$ จาก $f\left(\boldsymbol{\theta}_2 \middle| \boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_3^{(t-1)},...,\boldsymbol{\theta}_d^{(t-1)},\boldsymbol{y}\right)$

$\boldsymbol{\theta}_3^{(t)}$ จาก $f\left(\boldsymbol{\theta}_3 \middle| \boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_2^{(t)},\boldsymbol{\theta}_4^{(t-1)},...,\boldsymbol{\theta}_d^{(t-1)},\boldsymbol{y}\right)$

$\boldsymbol{\theta}_j^{(t)}$ จาก $f\left(\boldsymbol{\theta}_j \middle| \boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_2^{(t)},...,\boldsymbol{\theta}_{j-1}^{(t)},\boldsymbol{\theta}_{j+1}^{(t-1)},....,\boldsymbol{\theta}_d^{(t-1)},\boldsymbol{y}\right)$

$\vdots \qquad\qquad \vdots$

$\boldsymbol{\theta}_d^{(t)}$ จาก $f\left(\boldsymbol{\theta}_d \middle| \boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_2^{(t)},...,\boldsymbol{\theta}_{d-1}^{(t)},\boldsymbol{y}\right)$

ค่าพารามิเตอร์ที่ได้จากการสุ่มตัวอย่างแบบกิบส์สามารถสรุปได้ว่า

$$f\left(\boldsymbol{\theta}_j \middle| \boldsymbol{\theta}_1^{(t)},\boldsymbol{\theta}_2^{(t)},...,\boldsymbol{\theta}_{j-1}^{(t)},\boldsymbol{\theta}_{j+1}^{(t-1)},....,\boldsymbol{\theta}_d^{(t-1)},\boldsymbol{y}\right) \propto f\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$$

เมื่อได้ค่าพารามิเตอร์ที่ประมาณได้จากการสุ่มตัวอย่างแบบกิบส์ แล้วนำมาใช้ประมาณค่าพารามิเตอร์ในวิธีแบบเบส์ โดยประมาณค่าฟังก์ชันการแจกแจงภายหลัง ดังนี้

$$E\left(g\left(\theta\right)\middle|\boldsymbol{y}\right) = \int f\left(\theta\middle|y_1,y_2,...,y_n\right)g\left(\theta\right)d\theta \qquad (2.7)$$

ซึ่งสามารถประมาณค่า θ ได้ดังนี้

$$\bar{\boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)} \qquad (2.8)$$

2.6 วิธีการประมาณค่าข้อมูลสูญหาย

พิจารณาตัวอย่างขนาด n ที่สุ่มจากประชากรขนาด N ซึ่งประกอบด้วยค่าสังเกตของตัวแปรที่สนใจ Y และตัวแปรช่วย X ดังนี้ $(X_1,Y_1),...,(X_n,Y_n)$ สมมุติว่ามีหน่วยตัวอย่าง r หน่วยที่เก็บรวบรวมข้อมูลของตัวแปร Y ได้ $(r<n)$ นั่นคือ มีหน่วยตัวอย่าง $m=n-r$ หน่วย ที่ไม่สามารถเก็บรวบรวมข้อมูลของตัวแปร Y ได้ นั่นคือ $Y_{r+1},...,Y_n$ เป็นข้อมูลสูญหาย ในขณะที่ค่าของตัวแปร X สามารถเก็บรวบรวมได้จากทุกหน่วยตัวอย่าง นั่นคือข้อมูลของตัวแปร X เป็นข้อมูลที่เก็บรวบรวมได้ครบถ้วน ซึ่งแสดงได้ดังตารางที่ 1

ตาราง 1 ลักษณะของข้อมูลสูญหายที่เกิดขึ้นในตัวแปรตาม

X	Y	
x_1	y_1	
x_2	y_2	
x_3	y_3	
$\vdots$	$\vdots$	
x_r	y_r	
x_{r+1}	y_{r+1}	ข้อมูลสูญหาย
$\vdots$	$\vdots$	
x_n	y_n	

ในการศึกษานี้จะกล่าวถึงวิธีประมาณค่าสูญหายทั้งหมด 6 วิธี ดังนี้

2.6.1 วิธีสมการถดถอย (Regression Imputation: RI)

วิธีสมการถดถอยเป็นวิธีประมาณค่าสูญหายของตัวแปร Y โดยทำการประมาณสมการถดถอยของตัวแปร Y โดยใช้ข้อมูลที่ไม่สูญหาย (x_i, y_i) , $i = 1,...,r$ แล้วประมาณค่าข้อมูลสูญหายของตัวแปร Y โดยใช้สมการถดถอยที่หาได้ โดยกำหนดให้สมการถดถอยของตัวแปร Y ดังนี้

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} \qquad (2.9)$$

เมื่อ $\hat{y}_i$ แทน ค่าประมาณของตัวแปร y ของหน่วยตัวอย่างที่ i

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ แทน ค่าประมาณของสัมประสิทธิ์ของการถดถอย

X_i แทน ค่าของตัวแปรอิสระ x ที่ไม่สูญหายของหน่วยตัวอย่างที่ i , $i = 1,...,r$

โดยมีขั้นตอนการดำเนินงาน ดังนี้

<u>ขั้นตอนที่ 1</u> ใช้ข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ไม่สูญหายประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จากสูตร

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*\prime} \boldsymbol{X}^{*\prime} \right)^{-1} \boldsymbol{X}^{*\prime} \boldsymbol{Y}^* \qquad (2.10)$$

เมื่อ $\boldsymbol{X}^*$ และ $\boldsymbol{Y}^*$ แทนข้อมูลที่ไม่สูญหายของตัวแปร $\boldsymbol{X}$ และ $\boldsymbol{Y}$

ขั้นตอนที่ 2 เมื่อได้ค่า $\hat{\boldsymbol{\beta}}^*$ จากนั้นประมาณค่าสูญหายของตัวแปรตาม จากสมการถดถอยเชิงเส้นพหุคูณ

$$\begin{aligned}\hat{\boldsymbol{Y}}_i &= \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}^* \\ &= \begin{pmatrix}1 & \boldsymbol{X}_{i1} & \boldsymbol{X}_{i2} & \cdots & \boldsymbol{X}_{ip}\end{pmatrix}\hat{\boldsymbol{\beta}}^*\end{aligned} \qquad (2.11)$$

เมื่อ $\hat{\boldsymbol{Y}}_i$ แทน ค่าประมาณของค่าสังเกตชุดที่ i (ชุดที่ตัวแปรตามมีการสูญหาย)

$\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{ip}$ แทน ค่าสังเกตชุดที่ i ของตัวแปรอิสระตัวที่ $i = 1, 2, ..., p$

ขั้นตอนที่ 3 เมื่อได้ค่าของ $\hat{\boldsymbol{Y}}$ นำค่าของ $\hat{\boldsymbol{Y}}$ นั้น แทนในแถวที่มีข้อมูลสูญหาย จะได้ชุดข้อมูลที่สมบูรณ์เพื่อไปสร้างสมการถดถอยด้วยวิธีกำลังสองน้อยที่สุด

2.6.2 วิธีแทนค่าข้อมูลสูญหายหลายค่า (Multiple Imputation : MI)

วิธีนี้เป็นวิธีการประมาณค่าข้อมูลสูญหาย (Little and Rubin,1987) โดยจะประมาณค่าสูญหายให้กับข้อมูลที่มีการสูญหายหลายชุด ทำซ้ำเพื่อให้ได้ชุดข้อมูล u ชุด ในงานวิจัยนี้กำหนดให้ $u = 5$ มีขั้นตอนดังนี้ (สุปรียา สระโสม และธิดาเดียว มยุรีสวรรค์, 2547)

ขั้นตอนที่ 1 สร้างชุดข้อมูลสมบูรณ์จำนวน u ชุด โดยข้อมูลแต่ละชุดที่สร้างขึ้นมาใหม่กำหนดให้มีสถานการณ์เดียวกันกับข้อมูลในงานวิจัย เพื่อประมาณค่าพารามิเตอร์ $\hat{\boldsymbol{\beta}}$ ขึ้นมาใหม่ มีวิธีการดังนี้

1) นำข้อมูลตัวแปรตามและตัวแปรอิสระที่ไม่มีค่าสูญหายในตัวแปรตามมาสร้างตัวแบบถดถอยเชิงเส้นพหุคูณด้วยวิธี OLS ได้ตัวแบบถดถอยในรูป $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{obs}$ แล้วหาค่าความแปรปรวนของความคลาดเคลื่อนจากสูตร

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{m}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{obs})^2}{(m-p)} \qquad (2.12)$$

เมื่อ p แทน จำนวนสัมประสิทธิ์ถดถอย และ m แทนจำนวนข้อมูลที่ไม่สูญหายในตัวแปรตาม Y

2) สุ่มค่า a ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงไคกำลังสอง ด้วยองศาเสรีเท่ากับ $m-p$ เพื่อคำนวณค่า

$$\hat{\sigma}_*^2 = \frac{\hat{\sigma}_1^2(m-p)}{a} \qquad (2.13)$$

3) สุ่มค่า Z_j ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงปรกติมาตรฐานมาจำนวนเท่ากับสัมประสิทธิ์ถดถอย นำ Z_j มาคำนวณค่าสัมประสิทธิ์ถดถอยตัวที่ j ใหม่ จากสูตร

$$\hat{\beta}_j = \hat{\beta}_j^{obs} + \hat{\sigma}_* \left(\frac{S(\hat{\beta}_j^{obs})}{\hat{\sigma}_1} \right) Z_j \qquad (2.14)$$

เมื่อ $\hat{\beta}_j^{obs}$ และ $S(\hat{\beta}_j^{obs})$ แทน สัมประสิทธิ์ถดถอยและค่าความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอยตัวที่ j ตามลำดับ

4) สุ่มค่า Z ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงปรกติมาตรฐานมา 1 ค่า แล้วคำนวณค่าความคลาดเคลื่อนสุ่ม $Z\hat{\sigma}_*$ สำหรับใช้ปรับค่าสัมประสิทธิ์ถดถอย $\hat{\beta}_j$ ตัวใหม่ และนำตัวแบบถดถอยใหม่นี้ไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม Y ดังสมการต่อไปนี้

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} + Z\hat{\sigma}_* \qquad (2.15)$$

จะได้ข้อมูลชุดที่ 1 จากนั้นทำซ้ำ จนกระทั่งได้ชุดข้อมูลสมบูรณ์ทั้งหมด u ชุด

ขั้นตอนที่ 2 นำชุดข้อมูลสมบูรณ์แต่ละชุดมาสร้างตัวแบบถดถอยด้วยวิธี OLS จะได้ตัวแบบถดถอย u ตัวแบบ

ขั้นตอนที่ 3 จากตัวแบบถดถอย u ตัวแบบ ทำการหาค่าสัมประสิทธิ์ถดถอยใหม่จากค่าเฉลี่ยของสัมประสิทธิ์ถดถอยที่ตรงกัน เช่น ค่า $\hat{\beta}_0$ ใหม่คำนวณจาก $(\hat{\beta}_{01} + \hat{\beta}_{02} + \cdots + \hat{\beta}_{0u})/u$ แล้วนำค่าสัมประสิทธิ์ถดถอยใหม่ที่ได้ไปสร้างเป็นตัวแบบถดถอยเพื่อใช้ประมาณค่าสูญหายของตัวแปรตาม

2.6.3 วิธีค่าคาดหวังสูงสุด (Expectation Maximization Algorithm: EM)

วิธีนี้เป็นวิธีการประมาณค่าข้อมูลสูญหายที่หาค่าประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด โดยใช้กระบวนการวนซ้ำ แบ่งออกเป็น 2 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 E-Step (Expectation Step) เป็นขั้นตอนที่ใช้หาค่าคาดหวังของค่าสูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ไม่สูญหายและพารามิเตอร์ตัวปัจจุบัน นำค่าที่ได้นี้ไปประมาณค่าที่สูญหาย

ขั้นตอนที่ 2 M-Step (Maximization Step) เป็นขั้นตอนการประมาณค่าภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ด้วยการแทนค่าสูญหายที่ได้จากขั้นตอนที่ 1 จากนั้นทำซ้ำจนกระทั่งได้ตัวประมาณที่คงที่ จะได้ ตัวประมาณภาวะน่าจะเป็นสูงสุดโดย Little and Rubin, (1987) ได้นำวิธี EM มาประยุกต์ใช้ในการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยมีขั้นตอนดังนี้

1) สมมติมีข้อมูลดังนี้

$$\begin{bmatrix} \boldsymbol{Y}_{\mathbf{obs}} \\ \boldsymbol{Y}_{\mathbf{mis}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{\mathbf{obs}} \\ \boldsymbol{X}_{\mathbf{mis}} \end{bmatrix} \underset{\sim}{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$$

เมื่อ $\boldsymbol{Y}_{\mathbf{obs}}$ แทน เมทริกซ์ของตัวแปรตามที่ไม่สูญหาย

$\boldsymbol{Y}_{\mathbf{mis}}$ แทน เมทริกซ์ของตัวแปรตามที่สูญหาย

$\boldsymbol{X}_{\mathbf{obs}}$ แทน เมทริกซ์ของตัวแปรอิสระขนาดของชุดข้อมูลที่ตัวแปรตามไม่สูญหาย

$\boldsymbol{X}_{\mathbf{mis}}$ แทน เมทริกซ์ของตัวแปรอิสระขนาด ของชุดข้อมูลที่ตัวแปรตามสูญหาย

$\underset{\sim}{\boldsymbol{\beta}}$ แทน เมทริกซ์ของพารามิเตอร์

$\boldsymbol{\varepsilon}$ แทน เมทริกซ์ของค่าความคลาดเคลื่อน

2) ประมาณค่าเริ่มต้นของสัมประสิทธิ์ถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จากชุดข้อมูลที่สมบูรณ์ โดยจะเรียกว่า สัมประสิทธิ์ถดถอยรอบที่ t, $(t = 0,1,2,...,M)$

$$\hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t)} = \left(\boldsymbol{X}'_{\mathbf{obs}}\boldsymbol{X}_{\mathbf{obs}}\right)^{-1}\boldsymbol{X}'_{\mathbf{obs}}\boldsymbol{Y}_{\mathbf{obs}}$$

3) จาก $\hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t)}$ ที่ได้ในขั้นตอนที่ 2 สามารถหาค่าประมาณของข้อมูลสูญหายได้จากการหาค่าคาดหวัง ในขั้นตอน E-Step ได้ค่าคาดหวังรอบที่ t ได้ดังนี้

$$\left(y_i \mid \boldsymbol{Y}_{\mathbf{obs}}, \boldsymbol{X}_{\mathbf{obs}}, \hat{\underset{\sim}{\boldsymbol{\beta}}}^{(0)}\right) = \begin{cases} y_i; i = 1,2,\ldots,r \\ \hat{\beta}_0^{(0)} + \sum_{k=1}^{2} x_{ik}\hat{\beta}_k^{(0)}; i = r+1,\ldots,n \end{cases} \qquad (2.16)$$

ดังนั้น จะได้ $y_i^{(t)} = E\left(y_i \mid \boldsymbol{Y}_{\mathbf{obs}}, \boldsymbol{X}_{\mathbf{obs}}, \hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t-1)}\right)$

4) เมื่อหาค่าคาดหวังสำหรับทุกกรณีได้แล้ว จะประมาณค่าของพารามิเตอร์ด้วยการแทนค่าข้อมูลสูญหายที่ได้จาก E-Step เข้าสู่ขั้นตอน M-Step ในการทำซ้ำรอบที่ t

$$\hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

5) ถ้าหาก $\left|\hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t-1)} - \hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t)}\right| < 0.001$ จะได้ค่าประมาณของข้อมูลสูญหายรอบที่ t แต่ถ้า $\left|\hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t-1)} - \hat{\underset{\sim}{\boldsymbol{\beta}}}^{(t)}\right| > 0.001$ จะทำการประมาณรอบใหม่ โดยทำตามขั้นตอนที่ 3-4 ไปเรื่อยๆ โดยที่เราจะ

ใช้ $\hat{\underset{\sim}{\beta}}^{(t)}$ ประมาณข้อมูลสูญหายแทน $\hat{\underset{\sim}{\beta}}^{(t-1)}$ จนกระทั่งค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์ถดถอยน้อยกว่า 0.001 จึงจะได้ค่าประมาณของข้อมูลสูญหายที่แท้จริง

6) แทนค่าข้อมูลสูญหายด้วยค่าที่ประมาณได้ และทำการหาค่า MSE

2.6.4 วิธีแบบเบส์ (Bayes' Method)

การประมาณพารามิเตอร์ด้วยวิธีของเบส์ (Baye's method) พัฒนาโดยนักคณิตศาสตร์ชาวอังกฤษ ซึ่งการประมาณพารามิเตอร์ตามแนวคิดแบบดั้งเดิม (Classical approach) นั้น เริ่มจากการสุ่มตัวอย่างจากประชากรที่มีฟังก์ชันหนาแน่นของความน่าจะเป็น $f_X(x;\theta)$ และถือว่าพารามิเตอร์ θ เป็นค่าคงที่ที่ไม่ทราบค่า แต่ในแนวคิดแบบเบส์ จะนำเอาความเชื่อหรือความรู้เดิมเกี่ยวกับ θ มาใช้ประกอบการประมาณค่า θ ดังนั้นจึงถือว่า θ เป็นค่าของตัวแปรสุ่ม Θ ที่มีฟังก์ชันหนาแน่นของความน่าจะเป็นรูปแบบใดรูปแบบหนึ่ง (เกตุจันทร์ จำปาไชยศรี, 2559)

ให้ $X_1, X_2, ..., X_n$ เป็นตัวอย่างสุ่มมาจากประชากรที่มีฟังก์ชันหนาแน่นของความน่าจะเป็น $f_X(x;\theta) = f_{X|\Theta}(x|\theta)$ โดย $\theta \in \Theta$ และเรียก $f_{X|\Theta}(x|\theta)$ ว่า ฟังก์ชันหนาแน่นของความน่าจะเป็นแบบมีเงื่อนไข (Conditional density function) ของ X

ฟังก์ชันหนาแน่นของความน่าจะเป็นร่วมของ $X_1, X_2, ..., X_n$ เมื่อกำหนดให้ $\theta = \Theta$ ทำได้โดย

$$f_{X|\Theta}(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \qquad (2.17)$$

ให้ Θ มีฟังก์ชันหนาแน่นของความน่าจะเป็นก่อนเป็น $\pi(\theta)$

ฟังก์ชันหนาแน่นของความน่าจะเป็นร่วมของ $X_1, X_2, ..., X_n$ และ $\theta = \Theta$ คือ

$$f_{X|\Theta}(x_1, x_2, ..., x_n|\theta)\cdot\pi(\theta) = f(x_1,...x_n|\theta)\cdot\pi(\theta) = \prod_{i=1}^{n} f(x_i|\theta)\cdot\pi(\theta) \quad (2.18)$$

ดังนั้น ฟังก์ชันหนาแน่นของความน่าจะเป็นภายหลังของ Θ สามารถคำนวณได้จาก

$$\begin{aligned} h_{\Theta|X}(\theta|X) &= \frac{f(x_1, x_2, ..., x_n, \theta)}{f(x_1, x_2, ..., x_n)} \\ &= \frac{f(x_1, x_2, ..., x_n|\theta)\cdot\pi(\theta)}{f(x_1, x_2, ..., x_n)} \\ &= \frac{\prod_{i=1}^{n} f(x_i|\theta)\cdot\pi(\theta)}{f(x_1, x_2, ..., x_n)} \end{aligned} \qquad (2.19)$$

นั่นคือ

$$h_{\Theta|X}(\theta \mid X) = \begin{cases} \dfrac{\prod_{i=1}^{n} f(x_i \mid \theta)\cdot\pi(\theta)}{\int_{-\infty}^{\infty}\prod_{i=1}^{n} f(x_i \mid \theta)\cdot\pi(\theta)d\theta} & \text{เมื่อ } \theta \text{ เป็นตัวแปรสุ่มชนิดต่อเนื่อง} \\ \dfrac{\prod_{i=1}^{n} f(x_i \mid \theta)\cdot\pi(\theta)}{\sum_{\forall\theta}\prod_{i=1}^{n} f(x_i \mid \theta)\cdot\pi(\theta)} & \text{เมื่อ } \theta \text{ เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง} \end{cases} \quad (2.20)$$

ฟังก์ชันการแจกแจงภายหลัง $h_{\Theta|X}(\theta \mid X)$ เป็นฟังก์ชันที่ขึ้นอยู่กับค่าพารามิเตอร์ θ ดังนั้น สามารถแยกเป็นค่าคงที่ที่ไม่ขึ้นอยู่กับค่าพารามิเตอร์ θ ออกได้ จึงไม่จำเป็นต้องนำมาพิจารณา โดยสามารถเขียนใหม่ได้ดังนี้

$$h(\theta \mid X) \propto f(X \mid \theta)\pi(\theta) \quad (2.21)$$

2.6.4.1 การแจกแจงก่อนที่ให้สารสนเทศน้อยมาก (Noninformative prior distribution)

ให้ $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2)$

กำหนดฟังก์ชันการแจกแจงก่อนที่ให้สารสนเทศน้อยมาก ดังนี้

$$\pi(\boldsymbol{\beta}) = c \quad \text{เมื่อ } c \text{ เป็นค่าคงที่} \quad (2.22)$$

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad (2.23)$$

เราได้ ฟังก์ชันหนาแน่นของความน่าจะเป็นก่อนร่วม (joint conjugate prior distribution) ของ $\boldsymbol{\beta}, \sigma^2$ เป็นดังนี้

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\cdot\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad (2.24)$$

ฟังก์ชันภาวะน่าจะเป็น (Likelihood function)

$$\begin{aligned} L(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})\right] \\ &= (2\pi)^{-\frac{n}{2}}\cdot(\sigma^2)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})\right] \end{aligned}$$

$$\propto \quad \left(\sigma^2\right)^{-\frac{n}{2}} \cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right] \quad (2.25)$$

พิจารณา

$$\begin{aligned}
\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right) &= \boldsymbol{Y}'\boldsymbol{Y}-\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Y}\boldsymbol{X}'\boldsymbol{\beta}'+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}+2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}+2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right) \qquad (2.26)
\end{aligned}$$

เมื่อ $vS^2=\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$

$$v=n-k$$

$$S^2=\frac{1}{n-k}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$$

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

นั่นคือ

$$\begin{aligned}
L(\boldsymbol{Y}\mid\boldsymbol{\beta},\boldsymbol{X},\sigma^2) &= \left(\sigma^2\right)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right] \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}\left(vS^2\right)\right]\cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right] \\
&= \left(\sigma^2\right)^{-\frac{n}{2}}exp\left[-\frac{vS^2}{2\sigma^2}\right]\cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right] \quad (2.27)
\end{aligned}$$

เราได้

$$h\left(\boldsymbol{\beta},\sigma^2\mid\boldsymbol{X},\boldsymbol{Y}\right)=L(\boldsymbol{Y}\mid\boldsymbol{\beta},\boldsymbol{X},\sigma^2)\cdot\pi\left(\boldsymbol{\beta},\sigma^2\right)$$

$$= \left(\sigma^2\right)^{-\frac{n}{2}} exp\left[-\frac{vS^2}{2\sigma^2}\right] \cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right] \cdot \left(\sigma^2\right)^{-1}$$

$$= \left(\sigma^2\right)^{-\frac{n}{2}-1} exp\left[-\frac{vS^2}{2\sigma^2}\right] \cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right] \quad (2.28)$$

คำนวณฟังก์ชันหนาแน่นของความน่าจะเป็นภายหลัง (posterior distribution) ของ $\boldsymbol{\beta}$ ได้ดังนี้

$$h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \int_{-\infty}^{\infty} h\left(\boldsymbol{\beta},\sigma^2\middle|\mathrm{X,Y}\right)d\sigma^2$$

$$= \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}-1} exp\left[-\frac{vS^2}{2\sigma^2}\right] \cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right] d\sigma^2$$

$$= \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}-1} \cdot exp\left[-\frac{1}{2\sigma^2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right] d\sigma^2$$

$$= \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}-1} \cdot exp\left[-\frac{1}{\sigma^2}\cdot\frac{1}{2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right] d\sigma^2 \quad (2.29)$$

จัดสมการให้อยู่ในรูปของการแจกแจง Inverse gamma โดยพิจารณาจากคุณสมบัติต่อไปนี้

$$1 = \int_0^{\infty} \frac{q^p}{\Gamma(p)} x^{-p-1} exp\left[-\frac{q}{x}\right] dx$$

$$\frac{\Gamma(p)}{q^p} = \int_0^{\infty} x^{-p-1} exp\left[-\frac{q}{x}\right] dx \quad (2.30)$$

เมื่อ $\Gamma(p) = (p-1)!$ จะได้ว่า

$$\frac{\Gamma(p)}{q^p} = \frac{(p-1)!}{q^p} = (p-1)!q^{-p} \quad (2.31)$$

เราได้ว่า $p = \frac{n}{2},\ q = \frac{1}{2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)$

นั่นคือ $h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \frac{\Gamma(p)}{q^p}$

$$= \frac{(p-1)!}{q^p}$$

$$= (p-1)!q^{-p}$$

$$\propto q^{-p} = q^{-\frac{n}{2}} \quad (2.32)$$

เมื่อแทนค่า q ในสมการ (2.32)

$$h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \left[\frac{1}{2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right]^{-\frac{n}{2}}$$

$$= \left(\frac{1}{2}\right)^{-\frac{n}{2}}\cdot\left[vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^{-\frac{n}{2}}$$

$$\propto \left[vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^{-\frac{n}{2}}\cdot\left(\frac{1}{vS^2}\right)^{-\frac{n}{2}}$$

$$\propto \left[1+\frac{1}{vS^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^{-\left(\frac{v+k}{2}\right)} \qquad (2.33)$$

นั่นคือ $\boldsymbol{\beta} \sim mvt\left(\hat{\beta},\left(\frac{v}{v-2}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}S^2\right)$

เมื่อ mvt แทน การแจกแจงแบบทีหลายตัวแปร ด้วยองศาเสรีเท่ากับ v

ดังนั้น การแจกแจงภายหลังของ $\boldsymbol{\beta}$ มีการแจกแจงแบบทีหลายตัวแปร (Multivariate t-distribution) ด้วยเวกเตอร์ค่าเฉลี่ย $\hat{\beta}$ และเมทริกซ์ความแปรปรวนร่วม $\left(\frac{v}{v-2}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}S^2$

จากนั้น คำนวณฟังก์ชันหนาแน่นของความน่าจะเป็นภายหลัง (posterior distribution) ของ σ^2 ได้ดังนี้

$$h\left(\sigma^2\middle|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\beta}\right) = \int_{-\infty}^{\infty} h\left(\boldsymbol{\beta},\sigma^2\middle|\mathrm{X},\mathrm{Y}\right)d\boldsymbol{\beta}$$

$$= \int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\frac{n}{2}-1}exp\left[-\frac{vS^2}{2\sigma^2}\right]\cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]d\boldsymbol{\beta}$$

$$= \int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\frac{n}{2}-\frac{1}{2}}exp\left[-\frac{vS^2}{2\sigma^2}\right]\cdot\left(\sigma^2\right)^{-\frac{1}{2}}exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]d\boldsymbol{\beta}$$

$$= \left(\sigma^2\right)^{-\frac{n}{2}+\frac{1}{2}-1}exp\left[-\frac{vS^2}{2\sigma^2}\right]\cdot\int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\frac{1}{2}}exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]d\boldsymbol{\beta}$$

$$= \left(\sigma^2\right)^{-\left(\frac{n-1}{2}\right)-1}exp\left[-\frac{vS^2}{2\sigma^2}\right] \qquad (2.34)$$

ซึ่งเทอมปริพันธ์มีค่าเท่ากับ 1 เนื่องจากเป็นการหาปริพันธ์ตลอดช่วงของฟังก์ชันหนาแน่นของความน่าจะเป็นของการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ $\hat{\boldsymbol{\beta}}$ และความแปรปรวนเท่ากับ $\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$

และจะได้ว่า $\sigma^2 \sim Inv-gamma\left(\frac{n-1}{2},\frac{(n-k)S^2}{2}\right)$

จากนั้นใช้กระบวนการสุ่มตัวอย่างแบบกิบส์ (Gibbs sampling) ซึ่งเป็นวิธีการหนึ่งในวิธีมอนติคาร์โลลูกโซ่มาร์คอฟ (Markov chain monte carlo methods : MCMC) ประมาณค่าตัวแปรตามที่สูญหายเมื่อใช้การแจกแจงก่อนที่ให้สารสนเทศน้อยมาก โดยมีขั้นตอนดังนี้

1. กำหนด n แทน จำนวนข้อมูลทั้งหมด n_{obs} แทน จำนวนข้อมูลที่สังเกตได้ และ k แทน จำนวนสัมประสิทธิ์การถดถอย

2. จากชุดข้อมูลที่สังเกตได้ เราคำนวณค่า $\hat{\boldsymbol{\beta}}_{obs}=\begin{bmatrix}\beta_{0,obs}\\ \beta_{1,obs}\\ \beta_{2,obs}\end{bmatrix},\left(\boldsymbol{X}'_{obs}\boldsymbol{X}_{obs}\right)^{-1}$ และ

$S^2_{obs}=\frac{1}{n_{obs}-k}\left(\boldsymbol{Y}_{obs}-\boldsymbol{X}_{obs}\hat{\boldsymbol{\beta}}_{obs}\right)'\left(\boldsymbol{Y}_{obs}-\boldsymbol{X}_{obs}\hat{\boldsymbol{\beta}}_{obs}\right)$ และ $v_{obs}=n_{obs}-k$

3. สุ่ม $\boldsymbol{\beta}^{(0)} \sim mvt\left(\hat{\boldsymbol{\beta}}_{obs},\left(\frac{v_{obs}}{v_{obs}-2}\right)\left(\boldsymbol{X}'_{obs}\boldsymbol{X}_{obs}\right)^{-1}S^2_{obs}\right)$

$\sigma^{2(0)} \sim Inv-gamma\left(\frac{n_{obs}-1}{2},\frac{(n_{obs}-k)S^2_{obs}}{2}\right)$

4. เริ่มเข้าสู่กระบวนการสุ่มตัวอย่างแบบกิบส์ โดยสุ่มชุดข้อมูลสูญหายชุดที่ 1 จากพารามิเตอร์เริ่มต้น $\boldsymbol{\beta}^{(0)},\sigma^{2(0)}$ ที่ได้จากข้อที่ 3 และเรียกขั้นตอนนี้ว่า Imputation step (I-step) ได้ดังนี้ $\boldsymbol{Y}^{(1)}_{miss} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}^{(0)},\sigma^{2(0)}\right)$

5. จากข้อ 4 เมื่อได้ $\boldsymbol{Y}^{(1)}_{miss}$ แล้วนำ $\boldsymbol{Y}^{(1)}_{miss}$ มาประกอบกับชุดข้อมูลที่สังเกตได้ ซึ่งทำให้ได้ชุดข้อมูลที่สมบูรณ์ชุดที่ 1 $\boldsymbol{Y}^{(1)}=\begin{bmatrix}\boldsymbol{Y}_{obs}\\ \boldsymbol{Y}^{(1)}_{miss}\end{bmatrix}$ จากนั้นคำนวณค่า $\hat{\boldsymbol{\beta}}^{(1)}_{comp}=\begin{bmatrix}\beta_{0,comp}\\ \beta_{1,comp}\\ \beta_{2,comp}\end{bmatrix},\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ และ

$S^{2(1)}=\frac{1}{n-k}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(1)}_{comp}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(1)}_{comp}\right)$ สำหรับคำนวณค่าพารามิเตอร์ของการแจกแจงภายหลัง และเรียกขั้นตอนนี้ว่า Posterior step (P-step) ซึ่งได้ดังนี้

$\boldsymbol{\beta}^{(1)} \sim mvt\left(\hat{\boldsymbol{\beta}}^{(1)}_{comp},\left(\frac{v}{v-2}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}S^{2(1)}\right)$

$\sigma^{2(1)} \sim Inv-gamma\left(\frac{n-1}{2},\frac{(n-k)}{2}S^{2(1)}\right)$

6. (I-step) สุ่มชุดข้อมูลสูญหายชุดที่ 2 $\boldsymbol{Y}_{miss}^{(2)}$ โดยใช้ค่าพารามิเตอร์ในขั้นตอน P-step จากข้อ 5 ซึ่งได้ดังนี้ $\boldsymbol{Y}_{miss}^{(2)} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}^{(1)}, \sigma^{2(1)}\right)$ และคำนวณค่า $\hat{\boldsymbol{\beta}}_{comp}^{(2)} = \begin{bmatrix} \beta_{0,comp} \\ \beta_{1,comp} \\ \beta_{2,comp} \end{bmatrix}$, $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ และ $S^{2(2)} = \frac{1}{n-k}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{comp}^{(2)}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{comp}^{(2)}\right)$ จากชุดข้อมูลสมบูรณ์ชุดที่ 2

7. (P-step) คำนวณค่าพารามิเตอร์ของการแจกแจงภายหลัง

$$\boldsymbol{\beta}^{(2)} \sim mvt\left(\hat{\boldsymbol{\beta}}_{comp}^{(2)}, \left(\frac{v}{v-2}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} S^{2(2)}\right)$$

$$\sigma^{2(2)} \sim Inv-gamma\left(\frac{n-1}{2}, \frac{(n-k)}{2} S^{2(2)}\right)$$

8. ทำซ้ำขั้นตอนในข้อ 6 และ 7 จนกระทั่งได้ค่าพารามิเตอร์ของการแจกแจงภายหลังทั้งหมด t ค่า $\left(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right): t = 1, 2, ..., T$ ซึ่งจะได้ชุดข้อมูลสูญหายทั้งหมด t ชุด ในงานวิจัยนี้กำหนด $t = 6,000$ รอบ

9. จากข้อ 8 เราจะพิจารณาว่า $\left(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right): t = 1, 2, ..., T$ มีค่าคงที่ (Stationary distribution) ตั้งแต่รอบที่ $t = 1,000$ เป็นต้นไป

10. หาค่าเฉลี่ยของค่าพารามิเตอร์ของการแจกแจงภายหลังตั้งแต่รอบที่ 1000 เป็นต้นไป จะได้ $\boldsymbol{\beta}_{mean} = \frac{\sum_{t=1,000}^{6,000} \boldsymbol{\beta}^{(t)}}{5,000}$ และ $\sigma_{mean}^2 = \frac{\sum_{t=1,000}^{6,000} \sigma^{2(t)}}{5,000}$ เป็นพารามิเตอร์ที่ใช้ประมาณค่าข้อมูลสูญหายได้ดังนี้ $\boldsymbol{Y}_{miss} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}_{mean}, \sigma_{mean}^2\right)$

2.6.4.2 การแจกแจงภายหลังเมื่อใช้การแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ (Informative prior distribution)

กำหนดให้ $\boldsymbol{Y}$ เป็นเวกเตอร์สุ่มที่มีการแจกแจงปรกติด้วยค่าเฉลี่ย $\boldsymbol{X}\boldsymbol{\beta}$ และความแปรปรวน σ^2 สามารถเขียนได้ในรูป

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2)$$

กำหนดการแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ (Albert J., 2009) ดังนี้

$$\sigma^2 \sim Inv-gamma\left(\frac{v_0}{2},\frac{v_0\sigma_0^2}{2}\right)$$

$$\boldsymbol{\beta}\left|\sigma^2 \sim N\left(\boldsymbol{A},\boldsymbol{\Sigma}\right);\boldsymbol{\Sigma}=\sigma^2\mathbf{I}\right.$$

เมื่อ $\boldsymbol{A},\boldsymbol{\Sigma},v_0,\sigma_0^2$ แทน ไฮเปอร์พารามิเตอร์ที่ทราบค่า

พิจารณาฟังก์ชันหนาแน่นของความน่าจะเป็นก่อน (prior distribution) ของ $\boldsymbol{\beta}$ และ σ^2

$$\pi\left(\boldsymbol{\beta}\left|\sigma^2\right.\right) \propto \left(\sigma^2\right)^{-\frac{k}{2}} exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\right] \quad (2.35)$$

$$\pi\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-\left(\frac{v_0}{2}+1\right)} exp\left[-\frac{v_0\sigma_0^2}{2\sigma^2}\right] \quad (2.36)$$

เราได้ ฟังก์ชันหนาแน่นของความน่าจะเป็นก่อนร่วม (joint conjugate prior distribution) ของ $\boldsymbol{\beta},\sigma^2$ เป็นดังนี้

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\sigma^2\right) &= \pi\left(\boldsymbol{\beta}\left|\sigma^2\right.\right)\cdot\pi\left(\sigma^2\right)\\ &= \left(\sigma^2\right)^{-\frac{k}{2}-\left(\frac{v_0}{2}+1\right)}\cdot exp\left[-\frac{1}{2\sigma^2}\left(v_0\sigma_0^2+\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\right)\right] \quad (2.37)\end{aligned}$$

พิจารณาฟังก์ชันภาวะน่าจะเป็น

$$\begin{aligned}L(\boldsymbol{Y}\mid\boldsymbol{\beta},\boldsymbol{X},\sigma^2) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right]\\ &= \left(2\pi\right)^{-\frac{n}{2}}\cdot\left(\sigma^2\right)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right]\\ &\propto \left(\sigma^2\right)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right]\\ &= \left(\sigma^2\right)^{-\frac{n}{2}}\cdot exp\left[-\frac{1}{2\sigma^2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right] \quad (2.38)\end{aligned}$$

จะได้ ฟังก์ชันหนาแน่นของความน่าจะเป็นร่วม (joint probability density function) ของ $\boldsymbol{Y}$ มีรูปแบบดังนี้

$$\begin{aligned}h\left(\boldsymbol{\beta},\sigma^2\left|\boldsymbol{X},\boldsymbol{Y}\right.\right) &= L(\boldsymbol{Y}\mid\boldsymbol{\beta},\boldsymbol{X},\sigma^2)\cdot\pi\left(\boldsymbol{\beta},\sigma^2\right)\\ &= \left(\sigma^2\right)^{-\frac{n}{2}}exp\left[-\frac{1}{2\sigma^2}\left(vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)\right]\\ &\quad\times\left(\sigma^2\right)^{-\frac{k}{2}-\left(\frac{v_0}{2}+1\right)}exp\left[-\frac{1}{2\sigma^2}\left(v_0\sigma_0^2+\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\right)\right]\end{aligned}$$

$$= \left(\sigma^2\right)^{-\frac{n}{2}-\frac{k}{2}-\left(\frac{v_0}{2}+1\right)} exp\left[-\frac{1}{2\sigma^2}\begin{pmatrix} vS^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right) \\ +v_0\sigma_0^2+\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\end{pmatrix}\right]$$

$$= \left(\sigma^2\right)^{-\frac{n}{2}-\frac{k}{2}-\left(\frac{v_0}{2}+1\right)} exp\left[-\frac{1}{2\sigma^2}\begin{pmatrix} vS^2+v_0\sigma_0^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right) \\ +\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\end{pmatrix}\right]$$

พิจารณา

$$\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)$$

$$= \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$
$$+\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}-\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}-\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}+\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$$
$$= \boldsymbol{\beta}'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}-2\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}+\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$$
$$= \boldsymbol{\beta}'M\boldsymbol{\beta}-2\left[\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\beta}+\mathrm{K}$$
$$= \boldsymbol{\beta}'\mathrm{M}\boldsymbol{\beta}-2\mathrm{M}^{-1}\mathrm{M}\left[\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\beta}+\mathrm{K}$$
$$= \mathrm{M}\left[\boldsymbol{\beta}'\boldsymbol{\beta}-2\mathrm{M}^{-1}\left(\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\beta}\right]+\mathrm{K}$$
$$= \mathrm{M}\left[\boldsymbol{\beta}'\boldsymbol{\beta}-2\tilde{\mathrm{L}}\boldsymbol{\beta}\right]+\mathrm{K}$$
$$= \mathrm{M}\left[\boldsymbol{\beta}'\boldsymbol{\beta}-2\tilde{\mathrm{L}}\boldsymbol{\beta}\right]+\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}-\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}+\mathrm{K}$$
$$= \mathrm{M}\left[\boldsymbol{\beta}'\boldsymbol{\beta}-2\tilde{\mathrm{L}}\boldsymbol{\beta}+\tilde{\mathrm{L}}'\tilde{\mathrm{L}}\right]-\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}+\mathrm{K}$$
$$= \mathrm{M}\left[\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]-\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}+\mathrm{K}$$
$$= \left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)-\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}+\mathrm{K} \qquad (2.40)$$

เมื่อ $\mathrm{M} = \boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}$

$\tilde{\mathrm{L}} = \mathrm{M}^{-1}\left[\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\right]$

$\mathrm{K} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$

นั่นคือ

$$h\left(\boldsymbol{\beta},\sigma^2\middle|\boldsymbol{X},\boldsymbol{Y}\right) = \left(\sigma^2\right)^{-\frac{n}{2}-\frac{k}{2}-\left(\frac{v_0}{2}+1\right)} exp\left[-\frac{1}{2\sigma^2}\begin{pmatrix} vS^2+v_0\sigma_0^2 \\ +\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right) \\ -\mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}}+\mathrm{K}\end{pmatrix}\right]$$

$$= \left(\sigma^2\right)^{\frac{-\tilde{\nu}-k}{2}-1} exp\left[-\frac{1}{2\sigma^2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)\right] \quad (2.41)$$

เมื่อ $\tilde{\nu} = n+\nu_0$ และ $\tilde{S} = \nu S^2 + \nu_0\sigma_0^2 - \mathrm{M}\tilde{\mathrm{L}}'\tilde{\mathrm{L}} + \mathrm{K}$

คำนวณฟังก์ชันหนาแน่นของความน่าจะเป็นภายหลัง (posterior distribution) ของ $\boldsymbol{\beta}$ ได้ดังนี้

$$h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \int_{-\infty}^{\infty} h\left(\boldsymbol{\beta},\sigma^2\middle|\mathrm{X,Y}\right)d\sigma^2$$

$$= \int_{-\infty}^{\infty}\left(\sigma^2\right)^{\frac{-\tilde{\nu}-k}{2}-1} exp\left[-\frac{1}{2\sigma^2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)\right]d\sigma^2$$

$$= \int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)-1} exp\left[-\frac{1}{\sigma^2}\cdot\frac{1}{2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)\right]d\sigma^2$$

จัดสมการให้อยู่ในรูปของการแจกแจง Inverse gamma โดยพิจารณาจากคุณสมบัติต่อไปนี้

$$1 = \int_0^{\infty}\frac{q^p}{\Gamma(p)}x^{-p-1}exp\left[-\frac{q}{x}\right]dx$$

$$\frac{\Gamma(p)}{q^p} = \int_0^{\infty}x^{-p-1}exp\left[-\frac{q}{x}\right]dx \quad (2.43)$$

เมื่อ $\Gamma(p) = (p-1)!$ จะได้ว่า

$$\frac{\Gamma(p)}{q^p} = \frac{(p-1)!}{q^p} = (p-1)!q^{-p} \quad (2.44)$$

เราได้ว่า $p = \frac{\tilde{\nu}+k}{2},\ q = \frac{1}{2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)$

นั่นคือ $h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \frac{\Gamma(p)}{q^p}$

$$= \frac{(p-1)!}{q^p}$$

$$= (p-1)!q^{-p}$$

$$\propto q^{-p} = q^{-\left(\frac{\tilde{\nu}+k}{2}\right)} \quad (2.45)$$

เมื่อแทนค่า q ในสมการ (2.45)

$$h\left(\boldsymbol{\beta}\middle|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right) = \left[\frac{1}{2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)\right]^{-\left(\frac{\tilde{\nu}+k}{2}\right)}$$

$$= \left(\frac{1}{2}\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)}\cdot\left[\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]^{-\left(\frac{\tilde{\nu}+k}{2}\right)}$$

$$\propto \left[\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]^{-\left(\frac{\tilde{\nu}+k}{2}\right)}\cdot\left(\frac{1}{\tilde{\nu}\tilde{S}}\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)}$$

$$\propto \left[1+\frac{1}{\tilde{\nu}}\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\frac{\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]}{\tilde{S}}\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]^{-\left(\frac{\tilde{\nu}+k}{2}\right)} \qquad (2.46)$$

นั่นคือ $\boldsymbol{\beta}\sim mvt\left(\tilde{\mathrm{L}},\left(\frac{\tilde{\nu}}{\tilde{\nu}-2}\right)\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1}\tilde{S}\right)$

เมื่อ mvt แทน การแจกแจงแบบทีหลายตัวแปร ด้วยองศาเสรีเท่ากับ $\tilde{\nu}$

ดังนั้น การแจกแจงภายหลังของ $\boldsymbol{\beta}$ มีการแจกแจงแบบทีหลายตัวแปร (Multivariate t-distribution)ด้วยเวกเตอร์ค่าเฉลี่ย $\tilde{\mathrm{L}}$ และเมทริกซ์ความแปรปรวนร่วม $\left(\frac{\tilde{\nu}}{\tilde{\nu}-2}\right)\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1}\tilde{S}$

จากนั้น คำนวณฟังก์ชันหนาแน่นของความน่าจะเป็นภายหลัง (posterior distribution) ของ σ^2 ได้ดังนี้

$$h\left(\sigma^2\left|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\beta}\right.\right)=\int_{-\infty}^{\infty}h\left(\boldsymbol{\beta},\sigma^2\left|\mathrm{X,Y}\right.\right)d\boldsymbol{\beta}$$

$$=\int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)-1}exp\left[-\frac{1}{2\sigma^2}\left(\tilde{S}+\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right)\right]d\boldsymbol{\beta}$$

$$=\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)-\frac{1}{2}-\frac{1}{2}}exp\left(-\frac{\tilde{S}}{2\sigma^2}\right)$$

$$\cdot\int_{-\infty}^{\infty}exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]d\boldsymbol{\beta}$$

$$=\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)-\frac{1}{2}}exp\left(-\frac{\tilde{S}}{2\sigma^2}\right)$$

$$\times\int_{-\infty}^{\infty}\left(\sigma^2\right)^{-\frac{1}{2}}exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)'\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]\left(\boldsymbol{\beta}-\tilde{\mathrm{L}}\right)\right]d\boldsymbol{\beta}$$

$$=\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k}{2}\right)+\frac{1}{2}-1}exp\left(-\frac{\tilde{S}}{2\sigma^2}\right)$$

$$=\left(\sigma^2\right)^{-\left(\frac{\tilde{\nu}+k-1}{2}\right)-1}exp\left(-\frac{\tilde{S}}{2\sigma^2}\right) \qquad (2.47)$$

ซึ่งเทอมปริพันธ์มีค่าเท่ากับ 1 เนื่องจากเป็นการหาปริพันธ์ตลอดช่วงของฟังก์ชันหนาแน่นของความน่าจะเป็นของการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ $\tilde{\mathrm{L}}$ และความแปรปรวนเท่ากับ $\sigma^2\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]$

และจะได้ว่า $\sigma^2 \sim Inv-gamma\left(\frac{\tilde{\nu}+k-1}{2},\frac{\tilde{S}}{2}\right)$

จากนั้นใช้กระบวนการสุ่มตัวอย่างแบบกิบส์ (Gibbs sampling) ซึ่งเป็นวิธีการหนึ่งในวิธีมอนติคาร์โลลูกโซ่มาร์คอฟ (Markov chain monte carlo methods : MCMC) ประมาณค่าตัวแปรตามที่สูญหายเมื่อใช้การแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ โดยมีขั้นตอนดังนี้

1. กำหนดค่าไฮเปอร์พารามิเตอร์ $\boldsymbol{A},\boldsymbol{\Sigma},\nu_0,\sigma_0^2$ และ n แทน จำนวนข้อมูลทั้งหมด
2. สุ่มค่าพารามิเตอร์เริ่มต้น เพื่อใช้ประมาณค่า $\boldsymbol{Y}_{miss}$ จาก

$$\boldsymbol{\beta}^{(0)} \sim N\left(\boldsymbol{A},\boldsymbol{\Sigma}\right)$$

$$\sigma^{2(0)} \sim Inv-gamma\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)$$

3. เริ่มเข้าสู่กระบวนการสุ่มตัวอย่างแบบกิบส์ โดยสุ่มชุดข้อมูลสูญหายชุดที่ 1 จากค่าพารามิเตอร์เริ่มต้นที่ได้จากข้อที่ 2 และเรียกขั้นตอนนี้ว่า Imputation step (I-step) ได้ดังนี้

$\boldsymbol{Y}_{miss}^{(1)} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}^{(0)},\sigma^{2(0)}\right)$

4. จากข้อ 3 เมื่อได้ $\boldsymbol{Y}_{miss}^{(1)}$ แล้วนำ $\boldsymbol{Y}_{miss}^{(1)}$ มาประกอบกับชุดข้อมูลที่สังเกตได้ ซึ่งทำให้ได้ชุดข้อมูลที่สมบูรณ์ชุดที่ 1 $\boldsymbol{Y}^{(1)}=\begin{bmatrix}\boldsymbol{Y}_{obs}\\ \boldsymbol{Y}_{miss}^{(1)}\end{bmatrix}$ จากนั้นคำนวณค่า $\tilde{\mathrm{L}}^{(1)},\tilde{\nu},\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1(1)}$ และ $\tilde{S}^{(1)}$ สำหรับคำนวณค่าพารามิเตอร์ของการแจกแจงภายหลัง และเรียกขั้นตอนนี้ว่า Posterior step (P-step) ซึ่งได้ดังนี้

$$\boldsymbol{\beta}^{(1)} \sim mvt\left(\tilde{\mathrm{L}}^{(1)},\left(\frac{\tilde{\nu}}{\tilde{\nu}-2}\right)\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1(1)}\tilde{S}^{(1)}\right)$$

$$\sigma^{2(1)} \sim Inv-gamma\left(\frac{\tilde{\nu}+k-1}{2},\frac{\tilde{S}^{(1)}}{2}\right)$$

5. (I-step) สุ่มชุดข้อมูลสูญหายชุดที่ 2 $\boldsymbol{Y}_{miss}^{(2)}$ โดยใช้ค่าพารามิเตอร์จากข้อ 4 ได้ดังนี้ $\boldsymbol{Y}_{miss}^{(2)} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}^{(1)},\sigma^{2(1)}\right)$ และคำนวณค่า ค่า $\tilde{\mathrm{L}}^{(2)},\tilde{\nu},\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1(2)}$ และ $\tilde{S}^{(2)}$ จากชุดข้อมูลสมบูรณ์ชุดที่ 2

6. (P-step) คำนวณค่าพารามิเตอร์ของการแจกแจงภายหลัง

$$\boldsymbol{\beta}^{(2)} \sim mvt\left(\tilde{\mathrm{L}}^{(2)},\left(\frac{\tilde{\nu}}{\tilde{\nu}-2}\right)\left[\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}^{-1}\right]^{-1(2)}\tilde{S}^{(2)}\right)$$

$$\sigma^{2(2)} \sim Inv-gamma\left(\frac{\tilde{\nu}+k-1}{2}, \frac{\tilde{S}^{(2)}}{2}\right)$$

7. ทำซ้ำขั้นตอนในข้อ 5 และ 6 จนกระทั่งได้ค่าพารามิเตอร์ของการแจกแจงภายหลังทั้งหมด t ค่า $\left(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right): t = 1, 2, ..., T$ ซึ่งจะได้ชุดข้อมูลสูญหายทั้งหมด t ชุด ในงานวิจัยนี้กำหนด $t = 6,000$ รอบ

8. จากข้อ 7 เราจะพิจารณาว่า $\left(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right): t = 1, 2, ..., T$ มีค่าคงที่ (Stationary distribution) ตั้งแต่รอบที่ $t = 1,000$ เป็นต้นไป

9. หาค่าเฉลี่ยของค่าพารามิเตอร์ของการแจกแจงภายหลังตั้งแต่รอบที่ 1000 เป็นต้นไป จะได้ $\boldsymbol{\beta}_{mean} = \frac{\sum_{t=1,000}^{6,000} \boldsymbol{\beta}^{(t)}}{5,000}$ และ $\sigma^2_{mean} = \frac{\sum_{t=1,000}^{6,000} \sigma^{2(t)}}{5,000}$ เป็นพารามิเตอร์ที่ใช้ประมาณค่าข้อมูลสูญหาย ได้ดังนี้ $\boldsymbol{Y}_{miss} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}_{mean}, \sigma^2_{mean}\right)$

2.6.5 วิธีการถดถอยแบบเบส์บูตสแตรป (Bayes bootstrap regression Imputation : BBRI)

วิธีนี้เป็นการประมาณค่าพารามิเตอร์ด้วยวิธีเบส์บูตสแตรป (Donald B. Rubin., 1981) ในตัวแบบการถดถอยเพื่อประมาณค่าตัวแปรตามที่สูญหาย ซึ่งมีขั้นตอนดังนี้

1) กำหนดให้ r เป็นจำนวนข้อมูลที่สมบูรณ์ สุ่ม $\mathbf{U} = \left[u_1, u_2, ..., u_{r-1}\right]$ จากการแจกแจงเอกรูปที่อยู่ในช่วง 0 และ 1 มา $r-1$ ค่า แล้วเรียงลำดับจากน้อยไปมาก จะได้ $\mathbf{U}^* = \left[u_{(1)}, u_{(2)}, ..., u_{(r-1)}\right]$ เมื่อ $u_{(1)} < u_{(2)} < \cdots < u_{(r-1)}$ และคำนวณ $g_i = u_{(i)} - u_{(i-1)}$ เมื่อ $i = 1, 2, ..., r$ โดยที่ $u_{(0)} = 0$ และ $u_{(r)} = 1$ จะได้ $g = \left(g_1, g_2, ..., g_r\right)$ แทน เวกเตอร์ของความน่าจะเป็น โดยที่

$$g_1 = u_{(1)} - u_{(0)} = u_{(1)} - 0 = u_{(1)}$$
$$g_2 = u_{(2)} - u_{(1)}$$
$$\vdots$$
$$g_r = u_{(r)} - u_{(r-1)} = 1 - u_{(r-1)}$$

2) สุ่มตัวอย่างบูตสแตรป (Bootstrap sample) โดยวิธีการสุ่มแบบใส่คืน (sampling with replacement) จากชุดของข้อมูลที่สังเกตได้ จะได้ตัวอย่างบูตสแตรป ดังนี้

$$\left(y^B_{(1)}, x^B_{1(1)}, x^B_{2(1)}\right), \left(y^B_{(2)}, x^B_{1(2)}, x^B_{2(2)}\right), ..., \left(y^B_{(r)}, x^B_{1(r)}, x^B_{2(r)}\right)$$

และให้น้ำหนักข้อมูลด้วยความน่าจะเป็นที่ได้จากขั้นตอนที่ 1 ดังนี้

$$\left(g_1 y^B_{(1)}, g_1 x^B_{1(1)}, g_1 x^B_{2(1)}\right), \left(g_2 y^B_{(2)}, g_2 x^B_{1(2)}, g_2 x^B_{2(2)}\right), ..., \left(g_r y^B_{(r)}, g_r x^B_{1(r)}, g_r x^B_{2(r)}\right)$$

3) คำนวณค่าสัมประสิทธิ์ถดถอยของชุดข้อมูลด้วยวิธีกำลังสองน้อยที่สุด เราจะได้ $\beta^B_0, \beta^B_1, \beta^B_2$

4) ทำซ้ำขั้นตอนที่ 2-3 จำนวน 100 รอบ จะได้ $\beta^B_{0_k}, \beta^B_{1_k}$ และ $\beta^B_{2_k}$ ทั้งหมด 100 ค่า

5) คำนวณค่าสัมประสิทธิ์ถดถอยโดยเฉลี่ยที่ได้จากขั้นตอนที่ 4 นั่นคือ

$$\beta_{0B_j} = \frac{\sum_{k=1}^{100} \beta^B_{0_k}}{100}, \beta_{1B_j} = \frac{\sum_{k=1}^{100} \beta^B_{1_k}}{100} \text{ และ } \beta_{2B_j} = \frac{\sum_{k=1}^{100} \beta^B_{2_k}}{100}$$

6) ทำซ้ำขั้นตอนที่ 2-5 จำนวน 1,000 รอบ แล้วจะได้ $\beta_{0B_j}, \beta_{1B_j}, \beta_{2B_j}$ ทั้งหมด 1,000 ค่า แล้วหาค่าเฉลี่ย

$$\beta_{0BB} = \frac{\sum_{j=1}^{1,000} \beta_{0B_j}}{1,000}, \beta_{1BB} = \frac{\sum_{j=1}^{1,000} \beta_{1B_j}}{1,000} \text{ และ } \beta_{2BB} = \frac{\sum_{j=1}^{1,000} \beta_{2B_j}}{1,000}$$

7) ประมาณค่าข้อมูลสูญหายในตัวแปรตาม จากสมการ $\hat{y}^*_i = \beta_{0BB} + \beta_{1BB} x_{i1} + \beta_{2BB} x_{i2}$

2.7 เกณฑ์การตัดสินใจ

การเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายทั้ง 6 วิธี จะพิจารณาจากค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average mean square error: AMSE) สามารถคำนวณได้จากสูตร

$$AMSE = \frac{1}{10,000} \sum_{t=1}^{10,000} MSE_t$$

เมื่อ MSE_t แทน ค่า MSE ของแต่ละวิธี

b แทน จำนวนรอบของการทำซ้ำ เมื่อ $t = 1, 2, ..., 10,000$

2.8 งานวิจัยที่เกี่ยวข้อง

จริยา แสงสุวรรณ (2551) ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ 4 วิธีคือ วิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย และวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า (MI) สำหรับขนาดตัวอย่าง 50, 70, 100 และ 200 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1, 5 และ 15 เปอร์เซ็นต์การสูญหายของตัวแปรตามเท่ากับ 5%, 10%, 20% และ 30% ตามลำดับ และตัวแปรอิสระมีการแจกแจงแบบปรกติ ความสัมพันธ์ระหว่างตัวแปรอิสระมี 3 ระดับ คือ ระดับต่ำ (0.20) ระดับปานกลาง (0.50) และระดับสูง (0.70) ทำการจำลองด้วยวิธีมอนติคาร์โล ในแต่ละสถานการณ์กระทำซ้ำ 5,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบคือ ค่า RMSE โดยวิธีใดที่ให้ค่า RMSE ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่าวิธีสมการถดถอยและวิธีเอ็มไอให้ค่าประมาณของ RMSE ลดลงเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีการประมาณค่าสูญหายทั้ง 4 วิธี ให้ค่าประมาณของ RMSE แตกต่างกัน โดยวิธีสมการถดถอยและวิธีเอ็มไอ ให้ค่าประมาณของ RMSE ใกล้เคียงกัน

จิรกานต์ นวลละออง และเสาวณิต สุขภารังษี (2553) ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับตัวแบบการพยากรณ์ 3 วิธี คือ วิธีค่าเฉลี่ย วิธีกำลังสองน้อยที่สุด วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า และเสนอวิธีการใหม่เรียกว่า “การประมาณค่าสูญหายร่วม” โดยนำวิธีการประมาณค่าสูญหาย 3 วิธีมารวมกันด้วยตัวถ่วงน้ำหนัก ทำการจำลองข้อมูลด้วยวิธีมอนติคาร์โลด้วยโปรแกรม R ทำซ้ำ 50,000 รอบและกำหนดขนาดตัวอย่างเท่ากับ 30, 50 และ 100 ค่าเบี่ยงเบนมาตรฐานความคลาดเคลื่อนเท่ากับ 1, 2, 5 และ 10 เปอร์เซ็นต์ข้อมูลสูญหายเท่ากับ 5%, 10%, 15% และ 20% ขนาดความสัมพันธ์ของตัวแปรอิสระ คือ 0, 0.2 และ 0.5 เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ ผลสรุปที่ได้คือ วิธีที่เหมาะสมกับข้อมูลภาคตัดขวางคือ วิธีค่าเฉลี่ย วิธีที่เหมาะสมกับข้อมูลอนุกรมเวลาคือ วิธีกำลังสองน้อยที่สุด สำหรับวิธีการประมาณค่าร่วมกันนั้น วิธีถ่วงน้ำหนักโดยค่าสัมบูรณ์ต่ำสุดจะเหมาะสมทุกกรณี

เรืองลักษณ์ หลำใจซื่อ, อำไพ ทองธีรภาพ และจุฑาภรณ์ สินสมบูรณ์ทอง (2560) ศึกษาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายสาหรับการวิเคราะห์การถดถอยพหุเมื่อตัวแปรตามมีการสูญหายอย่างสุ่ม 4 วิธี ได้แก่ วิธี Regression Imputation (RI) วิธี Stochastic Regression Imputation (SRI) วิธี K-Nearest Neighbor (KNN) วิธี EM Algorithm (EM) และวิธีการประมาณค่าสูญหายแบบร่วม 2 วิธี ได้แก่ วิธี K-Nearest Regression Imputation with

Equivalent Weighted (KREW) และวิธี K-Nearest Stochastic Regression Imputation with Equivalent Weighted (KSEW) โดยใช้กับฟังก์ชันถ่วงน้ำหนักด้วยการให้น้ำหนักเท่ากัน (EW) จำลองข้อมูลด้วยวิธีมอนติคาร์โล กำหนดขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5, 10 และ 15 เปอร์เซ็นต์การสูญหาย 4 ระดับ คือ 10%, 20%, 30% และ 40% เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ผลการวิจัยพบว่า วิธี KSEW มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่าง 20 และ 30 วิธี SRI มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่าง 50 และ 100 และทุกวิธีจะมีประสิทธิภาพลดลงเมื่อเปอร์เซ็นต์การสูญหายและค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้น

สุปรียา สระโสม และ ธิดาเดียว มยุรีสวรรค์ (2562) ได้พัฒนาวิธีการแทนค่าข้อมูลสูญหายในตัวแปรตามสำหรับการถดถอยเชิงเส้นพหุคูณ เมื่อตัวแปรตามมีการสูญหายแบบสุ่ม (Missing at random) โดยวิธีที่พัฒนาได้แก่วิธี Mean Regression Imputation (MRI) วิธี Expectation Maximization with Multiple Imputation (EMMI) และวิธี Nearest Average Regression Imputation (NARI) โดยเปรียบเทียบประสิทธิภาพกับวิธีที่พัฒนาขึ้นกับอีก 6 วิธี ได้แก่วิธี Regression Imputation (RI) วิธี Stochastic Regression Imputation (SRI) วิธี K Nearest Neighbor Imputation (KNN) วิธี Expectation Maximization Algorithm (EM) วิธี Multiple Imputation (MI) และวิธี Proportioned Residual Draw Imputation (PRD) กำหนดส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน (σ) เท่ากับ 5, 10 และ 15 และขนาดตัวอย่าง (n) เท่ากับ 30, 50, 100 และ 200 และร้อยละการสูญหายเท่ากับ 5, 10, 15 และ 20 เกณฑ์เปรียบเทียบประสิทธิภาพคือค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean SquareError: AMSE) ผลการวิจัยพบว่า วิธี EMMI มีประสิทธิภาพดีที่สุดสำหรับทุกระดับขนาดตัวอย่างที่ σ มีค่าเท่ากับ 5 และร้อยละการสูญหายเท่ากับ 5 วิธี MRI มีประสิทธิภาพดีกว่าวิธีอื่นที่ทุกระดับขนาดตัวอย่างเมื่อ σ มีค่าเท่ากับ 10 และร้อยละการสูญหายเท่ากับ 5 และวิธี MRI ยังคงมีประสิทธิภาพดีที่สุดเมื่อ σ มีค่าเท่ากับ 15 ในทุกระดับร้อยละการสูญหายและเกือบทุกระดับขนาดตัวอย่าง ส่วนผลการศึกษาจากข้อมูลจริงที่ n เท่ากับ 50 พบว่าวิธี MRI มีประสิทธิภาพดีที่สุดในทุกระดับร้อยละการสูญหาย

Brandel (2004) ศึกษาเปรียบเทียบวิธีการแทนที่ค่าสูญหายด้วยวิธีการแบบเบส์ วิธีค่าคาดหวัง (EM) กับอีก 4 วิธี ได้แก่ วิธี Last Observation Carried Forward (LOCF) วิธีสมการถดถอย (Regression Imputation) วิธี Best or worst case imputation และวิธีค่าเฉลี่ย (Mean)

ผลการวิจัยพบว่า สำหรับทุกค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนและทุกระดับการสูญหาย โดยรวมแล้ววิธีการแบบเบส์มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่

Erlangung (2009) ศึกษาการวิเคราะห์ข้อมูลเชิงสำรวจที่มีค่าสูญหาย โดยได้ทำการศึกษาเปรียบเทียบทั้งหมด 4 วิธี ได้แก่ วิธี Bayesian Bootstrap Predictive Mean Matching (BBPMM) วิธี Posterior Predictive Mean Matching (PPMM) วิธี Rounded Predictive Mean Matching (RPMM) และวิธี Rounding to the nearest observed value (ROV) กำหนดขนาดตัวอย่างเป็นขนาดใหญ่และขนาดเล็กเท่ากับ 2,000 และ 200 ตามลำดับ กำหนดให้มีการสูญหายแบบ MAR และ MCAR เปอร์เซ็นต์การสูญหายของข้อมูลเป็น 60% โดยใช้ค่าความเอนเอียงสัมพัทธ์โดยเฉลี่ย (Average relative bias) และค่าความครอบคลุมเฉลี่ย (Average coverage) ที่ช่วงความเชื่อมั่น 95% เป็นเกณฑ์เปรียบเทียบประสิทธิภาพ ผลการวิจัยพบว่า ค่าความเอนเอียงเฉลี่ยของวิธี RPMM ดีกว่าวิธี ROV แต่ส่วนใหญ่ค่าครอบคลุมเฉลี่ยของทั้งสองวิธีมีค่าใกล้เคียงกัน กรณีชุดข้อมูลขนาดเล็กพบว่า วิธี PPMM และวิธี BBPMM มีค่าความเอนเอียงเฉลี่ยน้อยกว่าชุดข้อมูลขนาดใหญ่ ในขณะที่วิธี PPMM และ BBPMM ของชุดข้อมูลขนาดใหญ่ให้ค่าครอบคลุมเฉลี่ยที่สูงกว่า สำหรับรูปแบบการสูญหายแบบ MAR และ MCAR พบว่า วิธี PPMM และ BBPMM มีค่าความเอนเอียงเฉลี่ยน้อยที่สุด และให้ค่าครอบคลุมเฉลี่ยมากที่สุด โดยรวมแล้ว วิธี PPMM และ BBPMM มีค่าความเอนเอียงเฉลี่ยน้อยที่สุด แต่วิธี ROV มีค่าความเอนเอียงเฉลี่ยมากที่สุด เมื่อพิจารณาค่าครอบคลุมเฉลี่ย พบว่าส่วนใหญ่วิธี BBPMM มีค่าเข้าใกล้ 95% มากที่สุด และเป็นวิธีเดียวที่มีค่าความเชื่อมั่นครอบคลุมมากกว่า 90% ทุกกรณี

ธรรมรัตน์ กลีบเมฆ และนพคุณ ทองมวล (2563) ศึกษาการประมาณค่าข้อมูลสูญหายเมื่อตัวแปรตาม Y มีความสัมพันธ์กับตัวแปรอิสระ X โดยที่ตัวแปร X และ Y มีการแจกแจงปรกติ โดยเสนอวิธีประมาณค่าข้อมูลสูญหาย ด้วยวิธีการถดถอยแบบเบส์-บูตสแตรป และเปรียบเทียบกับวิธีการถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด โดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เป็นเกณฑ์เปรียบเทียบ เพื่อวัดความแม่นยำ ผลการศึกษาพบว่า เมื่อ n = 30, 50, 70, 90 และสัมประสิทธิ์สหสัมพันธ์เพิ่มขึ้น วิธีการถดถอยแบบเบย์-บูตสแตรปและวิธีการถดถอยมีความแม่นยำในการประมาณค่าสูญหายมากกว่าวิธีการถดถอยด้วยระยะทางต่ำสุด แต่เมื่อพิจารณาค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยในกรณี n = 30, 50 พบว่า วิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำกว่าวิธีการ

ถดถอยเล็กน้อย เมื่อขนาดตัวอย่างเพิ่มขึ้น พบว่าค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีมีค่าลดลง

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) วิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์ (Bay-in) วิธีแบบเบส์ที่ให้สาระสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI) ซึ่งรายละเอียดแบ่งออกเป็น 3 ส่วน ดังนี้

3.1 ขอบเขตงานวิจัย

3.2 ขั้นตอนการวิจัย

3.3 ขั้นตอนการทำงานของโปรแกรม

3.1 ขอบเขตงานวิจัย

3.1.1 กำหนดตัวแปรอิสร 2 ตัว มีการแจกแจงปรกติ นั่นคือ $X_1 \sim N(0,1)$ และ $X_2 \sim N(0,10)$ โดยที่ตัวแปรอิสระไม่มีการสูญหายและไม่มีความสัมพันธ์กัน

3.1.2 กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 0.5, 1, 2, 5, 10 และกำหนดขนาดตัวอย่างที่ศึกษาเป็น 50, 100 และ 200

3.1.3 สร้างตัวแปรตามให้มีความสัมพันธ์กับตัวแปรอิสระ โดยที่กำหนดระดับความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม 3 ระดับ คือ

1. ระดับต่ำ $0 \le \rho < 0.4$
2. ระดับปานกลาง $0.4 \le \rho < 0.7$
3. ระดับสูง $0.7 \le \rho < 1.0$

3.1.4 กำหนดให้ตัวแปรตามให้มีการสูญหายแบบสุ่ม (MAR) มีเปอร์เซ็นต์การสูญหายอยู่ที่ 5, 10 และ 20 ของขนาดตัวอย่างที่ศึกษา โดยกำหนดให้การสูญหายขึ้นอยู่กับตัวแปร X_2

3.1.5 กำหนดการแจกแจงก่อนที่ให้สารสนเทศน้อยมาก $f\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$

3.1.6 กำหนดการแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ $\boldsymbol{\beta} \sim N\left(\boldsymbol{A}, \boldsymbol{\Sigma}\right)$ และ $\sigma^2 \sim Inv-gamma\left(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}\right)$ เมื่อ $\boldsymbol{A}, \boldsymbol{\Sigma}, v_0, \sigma_0^2$ แทน ไฮเปอร์พารามิเตอร์ที่ทราบค่า

3.1.7 กำหนดจำนวนรอบบูตสแตรป เท่ากับ 1,000 รอบ

3.1.8 กำหนดจำนวนรอบของวิธีแทนค่าสูญหายหลายค่า (MI) จำนวน 5 รอบ

3.1.9 การศึกษานี้ได้ทำการจำลองค่าตัวแปรสุ่มตามการแจกแจงของประชากรที่กำหนด และทำซ้ำ 10,000 รอบ ในแต่ละสถานการณ์ โดยใช้โปรแกรม R studio ในการวิเคราะห์ข้อมูล

3.2 ขั้นตอนการวิจัย

3.2.1 กำหนดขนาดตัวอย่าง (n)

3.2.2 กำหนดความแปรปรวนของความคลาดเคลื่อน (σ^2) มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 0.5, 1, 2, 5 และ 10

3.2.3 สร้างตัวแปรอิสระ 2 ตัว มีการแจกแจงปรกติ นั่นคือ $X_1 \sim N(0,1)$ และ $X_2 \sim N(0,10)$ โดยที่ตัวแปรอิสระไม่มีการสูญหายและมีความสัมพันธ์กันต่ำ

3.2.4 สร้างตัวแปรตามให้มีความสัมพันธ์กับตัวแปรอิสระ โดยมีรูปแบบความสัมพันธ์เชิงเส้น ดังนี้

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{เมื่อ } i = 1, 2, ..., n$$

โดยกำหนดระดับความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม 3 ระดับ คือ

3.2.4.1 ระดับต่ำ $0 \le \rho < 0.4$

3.2.4.2 ระดับปานกลาง $0.4 \le \rho < 0.7$

3.2.4.3 ระดับสูง $0.7 \le \rho < 1.0$

3.2.5 สร้างตัวแปรตามให้มีการสูญหายแบบสุ่ม (MAR) มีเปอร์เซ็นต์การสูญหายอยู่ที่ 5, 10 และ 20 เมื่อเทียบกับขนาดตัวอย่างที่ศึกษา โดยกำหนดให้การสูญหายขึ้นอยู่กับตัวแปร X_2

3.2.6 ประมาณค่าข้อมูลของตัวแปรตามที่สูญหาย ด้วยวิธีประมาณทั้ง 6 วิธี

3.2.7 นำข้อมูลสมบูรณ์ที่ประมาณได้ของแต่ละวิธีไปสร้างตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares method : OLS) และหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE)

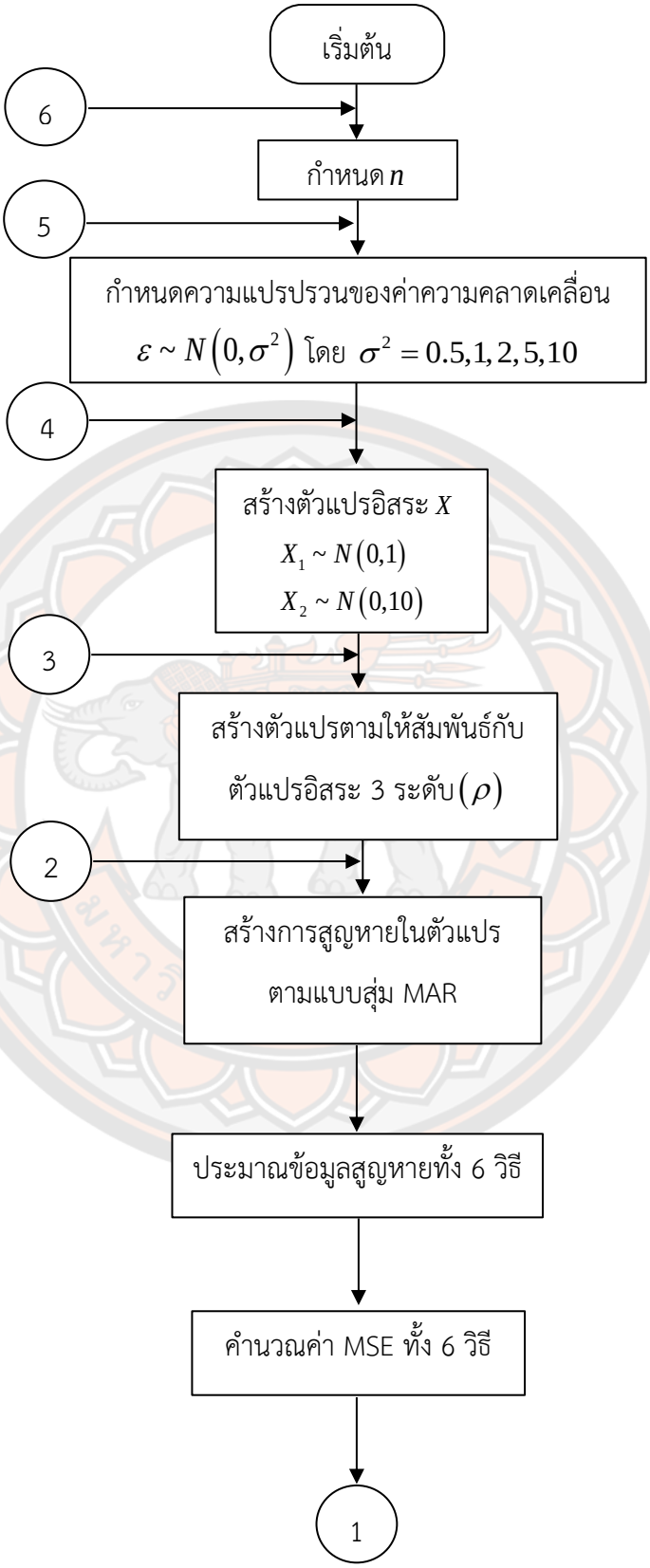
3.2.8 แต่ละสถานการณ์ทำซ้ำ 10,000 รอบ และหาค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) ของแต่ละวิธี

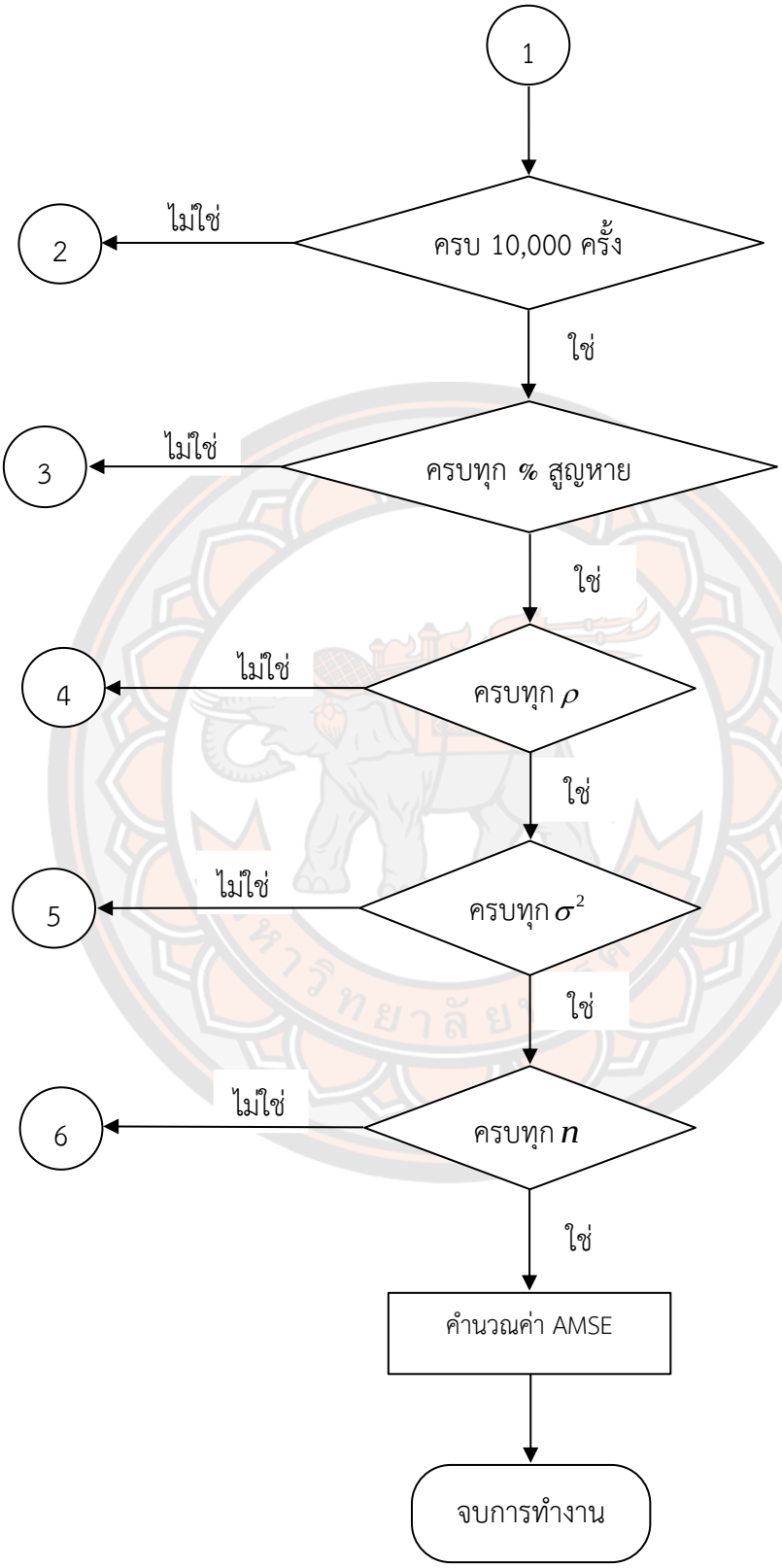
3.2.9 เปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) ของแต่ละวิธี โดยวิธีใดที่ให้ค่า AMSE ต่ำที่สุดจะเป็นวิธีที่มีประสิทธิภาพในการประมาณค่าสูญหายดีที่สุด

3.2.10 สรุปผลการวิจัย

3.3 ขั้นตอนการทำงานของโปรแกรม

ขั้นตอนในการทำงานของโปรแกรมสามารถเขียนให้อยู่ในรูปของผังงาน (Flowchart) ได้ดังนี้


เริ่มต้น
6
กำหนด n
5
กำหนดความแปรปรวนของค่าความคลาดเคลื่อน
$\varepsilon \sim N\left(0,\sigma^2\right)$ โดย $\sigma^2 = 0.5,1,2,5,10$
4
สร้างตัวแปรอิสระ X
$X_1 \sim N(0,1)$
$X_2 \sim N(0,10)$
3
สร้างตัวแปรตามให้สัมพันธ์กับ
ตัวแปรอิสระ 3 ระดับ (ρ)
2
สร้างการสูญหายในตัวแปร
ตามแบบสุ่ม MAR
ประมาณข้อมูลสูญหายทั้ง 6 วิธี
คำนวณค่า MSE ทั้ง 6 วิธี
1


1
ครบ 10,000 ครั้ง
ไม่ใช่
2
ใช่
ครบทุก % สูญหาย
ไม่ใช่
3
ใช่
ครบทุก ρ
ไม่ใช่
4
ใช่
ครบทุก σ^2
ไม่ใช่
5
ใช่
ครบทุก n
ไม่ใช่
6
ใช่
คำนวณค่า AMSE
จบการทำงาน

บทที่ 4

ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) วิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์ (Bay-in) วิธีแบบเบส์ที่ให้สาระสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI) กำหนดขนาดตัวอย่างเท่ากับ 50, 100 และ 200 เปอร์เซ็นต์การสูญหายของข้อมูลเท่ากับ 5, 10 และ 20 ความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5, 1, 2, 5, 10 ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ปานกลาง และสูง โดยใช้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) เป็นเกณฑ์ในการเปรียบเทียบ ซึ่งแบ่งหัวข้อได้ดังนี้

4.1 สัญลักษณ์ที่ใช้ในการวิจัย

4.2 ผลการวิจัย

4.1 สัญลักษณ์ที่ใช้ในการวิจัย

เพื่อให้เกิดความสะดวกในการอธิบายผลการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายแต่ละวิธีและมีความเข้าใจที่ถูกต้องตรงกัน ผู้วิจัยจึงได้กำหนดสัญลักษณ์ที่ใช้และความหมายดังต่อไปนี้

RI	แทน	วิธีสมการถดถอย
MI	แทน	วิธีแทนค่าข้อมูลสูญหายหลายค่า
EM	แทน	วิธีค่าคาดหวังสูงสุด
Bay-non	แทน	วิธีแบบเบส์ที่ให้สารสนเทศน้อยมาก
Bay-in	แทน	วิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์
BBRI	แทน	วิธีการถดถอยแบบเบส์บูตสแตรป
AMSE	แทน	ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ย
n	แทน	ขนาดตัวอย่าง
%	แทน	เปอร์เซ็นต์การสูญหายของข้อมูล
ρ	แทน	ระดับความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระ

4.2 ผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยจำแนกตามความแปรปรวนของค่าความคลาดเคลื่อน ได้ดังนี้

ตาราง 2 ค่า AMSE ของวิธีการประมาณข้อมูลสูญหาย 6 วิธี เมื่อจำแนกตามสถานการณ์ที่ศึกษา กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
ต่ำ	50	5	0.236	0.232	0.227	***0.222***	0.226	0.227
		10	0.238	0.214	0.211	***0.198***	0.209	0.201
		20	0.236	0.205	0.197	***0.179***	0.196	0.188
	100	5	0.243	0.231	0.229	0.237	0.235	***0.228***
		10	0.243	0.238	0.229	***0.220***	0.230	0.223
		20	0.245	0.244	0.232	***0.194***	0.229	0.230
	200	5	0.246	0.234	0.234	***0.215***	0.223	0.233
		10	0.246	0.232	0.221	***0.211***	0.226	0.221
		20	0.247	0.238	0.216	***0.205***	0.219	0.211
ปานกลาง	50	5	0.254	0.243	0.245	***0.229***	0.239	0.232
		10	0.253	0.227	0.227	***0.217***	0.229	0.219
		20	0.255	0.219	0.219	***0.210***	0.220	0.215
	100	5	0.256	0.243	0.242	***0.237***	0.240	0.243
		10	0.255	0.241	0.246	***0.230***	0.233	0.239
		20	0.256	0.235	0.233	***0.221***	0.230	0.224
	200	5	0.256	0.243	0.243	***0.233***	0.241	0.239
		10	0.256	0.230	0.231	***0.205***	0.224	0.232
		20	0.257	0.206	0.204	***0.199***	0.205	0.205
สูง	50	5	0.235	0.225	0.224	0.213	0.220	***0.211***
		10	0.235	0.219	0.219	0.218	0.225	***0.212***
		20	0.234	0.217	0.218	0.198	0.205	***0.187***
	100	5	0.243	0.231	0.231	***0.227***	0.234	0.230
		10	0.243	0.218	0.218	***0.193***	0.213	0.209
		20	0.242	0.224	0.222	0.209	0.225	***0.195***

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
สูง	200	5	0.246	0.234	0.234	0.225	0.244	***0.220***
		10	0.246	0.222	0.221	0.220	0.239	***0.199***
		20	0.247	0.227	0.196	0.193	0.198	***0.190***

หมายเหตุ : ตัวอักษรเอียง หมายถึง วิธีที่ให้ค่า AMSE ต่ำที่สุด ในแต่ละแถว

จากตารางที่ 2 กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5 พบว่า เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ที่ขนาดตัวอย่างเท่ากับ 50, 100 และ 200 พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำสุดเป็นส่วนใหญ่ในทุกเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ n =100 เปอร์เซ็นต์การสูญหายเท่ากับ 5 วิธี BBRI ให้ค่า AMSE ต่ำที่สุด เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับปานกลาง พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำที่สุดในทุกขนาดตัวอย่างและทุกเปอร์เซ็นต์การสูญหายของข้อมูล เมื่อความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามอยู่ในระดับสูง ที่ขนาดตัวอย่างเท่ากับ 50, 100 และ 200 พบว่าวิธี BBRI ให้ค่า AMSE ต่ำสุดเป็นส่วนใหญ่ในทุกเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ n =100 และเปอร์เซ็นต์การสูญหายเท่ากับ 5 และ 10 วิธี Bay-in ให้ค่า AMSE ต่ำที่สุด

ตาราง 3 ค่า AMSE ของวิธีการประมาณข้อมูลสูญหาย 6 วิธี จำแนกตามสถานการณ์ที่ศึกษา กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1

ρ	n	%	วิธีการประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
ต่ำ	50	5	0.944	0.924	0.919	0.903	0.914	***0.884***
		10	0.939	0.852	0.842	0.839	0.876	***0.837***
		20	0.946	0.753	0.744	0.754	0.771	***0.740***
	100	5	0.969	0.920	0.921	***0.906***	0.932	0.918
		10	0.971	0.889	0.892	0.879	0.897	***0.871***
		20	0.972	0.777	0.772	0.764	0.771	***0.756***
	200	5	0.983	0.964	0.934	***0.932***	0.950	0.937
		10	0.984	0.886	0.884	***0.878***	0.884	0.883
		20	0.986	0.792	0.785	***0.737***	0.789	0.788
ปานกลาง	50	5	0.911	0.873	0.874	***0.860***	0.884	0.873
		10	0.909	0.814	0.812	***0.791***	0.820	0.815
		20	0.911	0.824	0.815	***0.778***	0.809	0.787

ρ	n	%	วิธีการประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
ปานกลาง	100	5	0.952	0.929	0.924	0.909	0.919	***0.904***
		10	0.950	0.856	0.857	0.856	0.866	***0.854***
		20	0.949	0.759	0.755	0.745	0.754	***0.740***
	200	5	0.973	0.954	***0.924***	0.932	0.945	***0.924***
		10	0.976	0.880	0.885	0.879	***0.878***	***0.878***
		20	0.972	0.780	0.781	0.776	0.770	***0.764***
สูง	50	5	0.938	0.921	0.920	0.903	0.919	***0.899***
		10	0.939	0.876	***0.843***	0.879	0.856	***0.843***
		20	0.944	0.745	0.741	0.739	0.736	***0.730***
	100	5	0.966	0.949	***0.932***	0.935	0.942	***0.932***
		10	0.970	0.883	0.872	0.870	0.876	***0.868***
		20	0.968	0.775	0.771	0.772	0.774	***0.764***
	200	5	0.984	0.947	0.938	***0.923***	0.940	0.938
		10	0.983	0.886	0.885	***0.866***	0.889	0.884
		20	0.973	0.792	0.787	***0.776***	0.786	0.787

หมายเหตุ : ตัวอักษรเอียง หมายถึง วิธีที่ให้ค่า AMSE ต่ำที่สุด ในแต่ละแถว

จากตารางที่ 3 กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1 พบว่า เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ที่ขนาดตัวอย่างเท่ากับ 50 พบว่าวิธี BBRI ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล และที่ขนาดตัวอย่างเท่ากับ 200 พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ n =100 เปอร์เซ็นต์การสูญหายเท่ากับ 5 วิธี Bay-in ให้ค่า AMSE ต่ำที่สุด และที่เปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 วิธี Boot-Bay ให้ค่า AMSE ต่ำที่สุด

เมื่อความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามอยู่ในระดับปานกลาง ที่ขนาดตัวอย่างเท่ากับ 50 พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล และที่ขนาดตัวอย่างเท่ากับ 100 พบว่าวิธี BBRI ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล ส่วนที่ขนาดตัวอย่างเท่ากับ 200 วิธี BBRI ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหาย แต่ในเปอร์เซ็นต์การสูญหายเท่ากับ 5 และ 10 วิธี EM และวิธี Bay-in ให้ค่า AMSE ต่ำที่สุดใกล้เคียงกับวิธี BBRI ตามลำดับ

เมื่อความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามอยู่ในระดับสูง ที่ขนาดตัวอย่างเท่ากับ 200 พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล และที่ขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี BBRI ให้ค่า AMSE ต่ำสุดในทุกเปอร์เซ็นต์การสูญหายของข้อมูล แต่ที่ n =50 เปอร์เซ็นต์การสูญหายเท่ากับ 10 และที่ n =100 เปอร์เซ็นต์การสูญหายเท่ากับ 5 วิธี EM ให้ค่า AMSE ต่ำที่สุดใกล้เคียงกับวิธี BBRI

ตาราง 4 ค่า AMSE ของวิธีการประมาณข้อมูลสูญหาย 6 วิธี จำแนกตามสถานการณ์ที่ศึกษา กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 2

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
ต่ำ	50	5	3.805	3.711	3.669	***3.646***	3.663	3.655
		10	3.802	3.564	3.403	***3.384***	3.395	3.406
		20	3.835	3.348	3.346	3.339	3.344	***3.312***
	100	5	3.892	3.691	3.694	***3.650***	3.673	3.661
		10	3.899	3.503	3.497	***3.475***	3.483	3.481
		20	3.903	3.119	3.099	***3.049***	3.088	3.093
	200	5	3.947	3.749	3.744	3.740	3.745	***3.739***
		10	3.933	3.545	3.542	***3.541***	***3.541***	3.547
		20	3.941	3.159	3.149	***3.140***	3.153	3.151
ปานกลาง	50	5	3.692	3.671	3.669	3.667	3.670	***3.652***
		10	3.826	3.453	3.437	***3.435***	3.447	3.436
		20	3.839	3.247	3.218	3.201	3.221	***3.155***
	100	5	3.908	3.739	3.723	3.723	3.732	***3.721***
		10	3.907	3.510	3.509	***3.499***	3.505	3.506
		20	3.899	3.113	3.099	***3.083***	3.119	3.108
	200	5	3.934	3.754	***3.749***	3.756	3.754	***3.749***
		10	3.929	3.545	3.539	3.539	3.539	***3.536***
		20	3.949	3.171	3.162	***3.146***	3.151	3.150
สูง	50	5	3.770	3.609	3.598	***3.577***	3.587	3.599
		10	3.771	3.371	3.369	***3.360***	3.370	3.363
		20	3.774	3.013	***2.967***	2.979	2.989	2.985
	100	5	3.873	3.681	3.687	3.676	3.687	***3.666***
		10	3.877	3.587	3.584	3.557	3.671	***3.493***
		20	3.878	3.101	3.080	3.086	3.070	***3.066***

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
		5	3.938	3.741	3.739	3.743	3.741	***3.722***
	200	**10**	3.939	3.547	3.547	3.541	3.541	***3.534***
		20	3.950	3.173	3.143	3.137	3.154	***3.135***

หมายเหตุ : ตัวอักษรเอียง หมายถึง วิธีที่ให้ค่า AMSE ต่ำที่สุด เรียงตามแถว

จากตารางที่ 4 กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 2 พบว่า เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ที่ขนาดตัวอย่างเท่ากับ 50, 100 และ 200 พบว่าวิธี Bay-in ให้ค่า AMSE ต่ำสุดเป็นส่วนใหญ่ในทุกเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ n =50 เปอร์เซ็นต์การสูญหายเป็น 20 และ n =200 เปอร์เซ็นต์การสูญหายเป็น 5 วิธี BBRI ให้ค่า AMSE ต่ำที่สุด แต่ที่ n =200 เปอร์เซ็นต์การสูญหายเท่ากับ 10 วิธี Bay-non ให้ค่า AMSE ต่ำที่สุดใกล้เคียงกับวิธี Bay-in

เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับปานกลาง ที่ขนาดตัวอย่างเท่ากับ 50 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุด ยกเว้นที่เปอร์เซ็นต์การสูญหายเป็น 10 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 100 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุด ยกเว้นที่เปอร์เซ็นต์การสูญหายเป็น 5 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 200 เปอร์เซ็นต์การสูญหายเป็น 5 วิธี BBRI และวิธี EM ให้ค่าประมาณ AMSE ต่ำที่สุดใกล้เคียงกัน ที่เปอร์เซ็นต์การสูญหายเป็น 20 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุด และที่เปอร์เซ็นต์การสูญหายเป็น 10 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุด

เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง ที่ขนาดตัวอย่างเท่ากับ 100 และ 200 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล แต่ที่ขนาดตัวอย่างเท่ากับ 50 วิธี Bay-in ให้ค่า AMSE ต่ำสุดเป็นส่วนใหญ่ ยกเว้นเปอร์เซ็นต์การสูญหายเป็น 20 วิธี EM ให้ค่า AMSE ต่ำที่สุด

ตาราง 5 ค่า AMSE ของวิธีการประมาณข้อมูลสูญหาย 6 วิธี จำแนกตามสถานการณ์ที่ศึกษา กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5

			วิธีประมาณค่าข้อมูลสูญหาย					
ρ	n	%	RI	MI	EM	Bay-in	Bay-non	BBRI
ต่ำ	50	5	23.503	23.040	22.582	22.574	22.592	***22.572***
		10	23.599	21.139	21.155	21.107	21.110	***21.097***
		20	23.590	18.747	18.760	18.758	18.754	***18.742***
	100	5	24.239	23.987	23.815	23.761	23.753	***22.990***
		10	24.256	21.818	21.741	21.749	21.754	***21.731***
		20	24.241	19.379	19.387	19.366	19.370	***19.354***
	200	5	24.589	23.382	23.379	23.339	23.342	***23.319***
		10	24.623	22.177	22.189	22.171	22.168	***22.135***
		20	24.690	19.823	19.684	19.668	19.670	***19.637***
ปานกลาง	50	5	24.211	23.489	23.431	23.284	23.297	***23.104***
		10	24.175	21.771	21.708	21.671	21.696	***21.632***
		20	24.205	19.178	19.184	18.813	18.923	***18.691***
	100	5	24.553	23.302	23.343	***21.681***	21.990	22.191
		10	24.609	22.097	21.994	***21.492***	21.590	22.013
		20	24.604	20.702	19.902	***19.330***	19.453	19.584
	200	5	24.747	23.506	23.469	23.490	23.496	***23.456***
		10	24.723	22.576	22.347	22.276	22.320	***22.238***
		20	24.736	19.849	19.805	19.772	19.765	***19.701***
สูง	50	5	23.399	22.402	22.534	22.221	22.323	***20.984***
		10	23.505	21.969	21.971	21.591	21.762	***20.975***
		20	23.481	19.614	19.613	19.043	19.165	***18.930***
	100	5	24.155	23.909	23.123	23.196	23.208	***23.029***
		10	24.179	21.761	21.710	21.697	21.717	***21.294***
		20	24.256	20.903	19.411	19.395	19.401	***19.347***
	200	5	24.648	23.408	23.323	***21.527***	21.930	23.343
		10	24.592	22.134	22.112	***21.389***	21.478	22.141
		20	24.616	19.768	19.647	***19.557***	19.611	19.637

หมายเหตุ : ตัวอักษรเอียง หมายถึง วิธีที่ให้ค่า AMSE ต่ำที่สุด เรียงตามแถว

จากตารางที่ 5 กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5 พบว่า เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุด ในทุกเปอร์เซ็นต์การสูญหายของข้อมูลและทุกขนาดตัวอย่าง เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับปานกลาง เมื่อขนาดตัวอย่างเท่ากับ 50 และ 200 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล และที่ขนาดตัวอย่างเท่ากับ 100 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง ที่ขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล และที่ขนาดตัวอย่างเท่ากับ 200 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล

ตาราง 6 ค่า AMSE ของวิธีการประมาณข้อมูลสูญหาย 6 วิธี เมื่อจำแนกตามสถานการณ์ที่ศึกษา กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 10

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
ต่ำ	50	5	94.589	92.547	91.836	91.581	91.632	***85.161***
		10	94.737	85.067	84.974	84.706	84.851	***83.526***
		20	94.832	75.277	74.966	74.007	74.103	***72.473***
	100	5	97.031	92.084	91.996	***91.767***	91.793	91.982
		10	97.122	87.289	87.448	***87.085***	87.135	87.680
		20	97.022	77.538	77.183	***76.042***	76.097	77.587
	200	5	98.450	93.711	93.850	93.652	93.694	***93.573***
		10	98.514	89.915	89.920	88.869	88.890	***88.740***
		20	98.658	79.097	78.772	78.734	78.758	***78.544***
ปานกลาง	50	5	97.152	93.105	93.024	92.841	93.001	***89.879***
		10	96.750	88.509	88.439	87.771	87.862	***86.556***
		20	96.769	76.871	75.965	75.529	75.849	***74.800***
	100	5	98.193	93.233	93.091	93.077	93.086	***93.025***
		10	98.120	89.192	89.115	89.080	89.121	***88.851***
		20	98.579	78.878	78.196	78.526	78.371	***78.099***
	200	5	98.590	94.642	93.947	93.801	93.997	***93.655***
		10	98.603	88.857	88.942	88.711	88.736	***88.095***
		20	98.710	79.236	78.611	78.339	78.440	***78.164***

ρ	n	%	วิธีประมาณค่าข้อมูลสูญหาย					
			RI	MI	EM	Bay-in	Bay-non	BBRI
สูง	50	5	94.668	90.679	89.768	90.919	91.009	***84.875***
		10	93.895	84.067	83.904	83.811	83.826	***83.226***
		20	93.862	74.551	74.401	74.347	74.382	***74.290***
	100	5	97.309	92.369	91.833	90.984	91.087	***90.224***
		10	97.267	87.621	87.438	87.386	87.410	***87.308***
		20	97.275	77.853	77.827	77.642	77.751	***76.910***
	200	5	98.517	94.570	94.287	93.882	93.992	***93.372***
		10	98.526	89.809	89.393	89.086	89.097	***88.614***
		20	98.486	79.125	78.300	78.245	78.332	***78.152***

หมายเหตุ : ตัวอักษรเอียง หมายถึง วิธีที่ให้ค่า AMSE ต่ำที่สุด เรียงตามแถว

จากตารางที่ 6 กรณีความแปรปรวนของความคลาดเคลื่อนเท่ากับ 10 พบว่า เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ที่ขนาดตัวอย่างเท่ากับ 50 และ 200 วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ขนาดตัวอย่างเท่ากับ 100 วิธี Bay-in ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล

เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับปานกลางและระดับสูง วิธี BBRI ให้ค่าประมาณ AMSE ต่ำที่สุดทุกเปอร์เซ็นต์การสูญหายของข้อมูล และทุกขนาดตัวอย่าง นอกจากนี้ยังพบว่า ที่ขนาดตัวอย่างเท่ากับ 50 สำหรับทุกระดับความสัมพันธ์ของตัวแปรอิสระและตัวแปรตามที่เปอร์เซ็นต์การสูญหายเป็น 5 พบว่าค่าประมาณ AMSE ของวิธี BBRI มีค่าต่ำที่สุดแตกต่างจากวิธีอื่นอย่างเห็นได้ชัด

ตาราง 7 สรุปวิธีประมาณค่าตัวแปรตามที่สูญหายที่มีค่า AMSE ต่ำที่สุดในแต่ละสถานการณ์

ρ	n	%	ความแปรปรวนของความคลาดเคลื่อน (σ^2)				
			0.5	1	2	5	10
ต่ำ	50	5	Bay-in	BBRI	Bay-in	BBRI	BBRI
		10	Bay-in	BBRI	Bay-in	BBRI	BBRI
		20	Bay-in	BBRI	BBRI	BBRI	BBRI
	100	5	BBRI	Bay-in	Bay-in	BBRI	Bay-in
		10	Bay-in	BBRI	Bay-in	BBRI	Bay-in
		20	Bay-in	BBRI	Bay-in	BBRI	Bay-in
	200	5	Bay-in	Bay-in	BBRI	BBRI	BBRI
		10	Bay-in	Bay-in	Bay-in, Bay-non	BBRI	BBRI
		20	Bay-in	Bay-in	Bay-in	BBRI	BBRI
ปานกลาง	50	5	Bay-in	Bay-in	BBRI	BBRI	BBRI
		10	Bay-in	Bay-in	Bay-in	BBRI	BBRI
		20	Bay-in	Bay-in	BBRI	BBRI	BBRI
	100	5	Bay-in	BBRI	BBRI	Bay-in	BBRI
		10	Bay-in	BBRI	Bay-in	Bay-in	BBRI
		20	Bay-in	BBRI	Bay-in	Bay-in	BBRI
	200	5	Bay-in	EM, BBRI	EM, BBRI	BBRI	BBRI
		10	Bay-in	Bay-non, BBRI	BBRI	BBRI	BBRI
		20	Bay-in	BBRI	Bay-in	BBRI	BBRI
สูง	50	5	BBRI	BBRI	Bay-in	BBRI	BBRI
		10	BBRI	EM, BBRI	Bay-in	BBRI	BBRI
		20	BBRI	BBRI	EM	BBRI	BBRI
	100	5	Bay-in	EM, BBRI	BBRI	BBRI	BBRI
		10	Bay-in	BBRI	BBRI	BBRI	BBRI
		20	BBRI	BBRI	BBRI	BBRI	BBRI
	200	5	BBRI	Bay-in	BBRI	Bay-in	BBRI
		10	BBRI	Bay-in	BBRI	Bay-in	BBRI
		20	BBRI	Bay-in	BBRI	Bay-in	BBRI

จากตารางที่ 7 กรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำและปานกลาง เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5 พบว่า วิธี Bay-in มีประสิทธิภาพดีที่สุด

เป็นส่วนใหญ่ และเมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1 และ 2 วิธี Bay-in และวิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ แต่เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5 และ 10 พบว่า วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกระดับขนาดตัวอย่าง และเปอร์เซ็นต์การสูญหายของข้อมูล ในกรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกค่าความแปรปรวนของความคลาดเคลื่อน

บทที่ 5

บทสรุป

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ได้แก่ วิธีสมการถดถอย (RI) วิธีแทนค่าข้อมูลสูญหายหลายค่า (MI) วิธีค่าคาดหวังสูงสุด (EM) วิธีแบบเบส์ที่ให้สารสนเทศที่เป็นประโยชน์ (Bay-in) วิธีแบบเบส์ที่ให้สาระสนเทศน้อยมาก (Bay-non) และวิธีการถดถอยแบบเบส์บูตสแตรป (BBRI) กำหนดขนาดตัวอย่างเท่ากับ 50, 100 และ 200 เปอร์เซ็นต์การสูญหายของข้อมูลเท่ากับ 5, 10 และ 20 ความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5, 1, 2, 5, 10 ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำ ปานกลาง และสูง โดยใช้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) เป็นเกณฑ์ในการเปรียบเทียบ

5.1 สรุปผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จำนวน 6 วิธี ซึ่งวิธีที่ให้ค่าประมาณ AMSE ต่ำที่สุด แสดงว่าเป็นวิธีที่มีประสิทธิภาพมากที่สุด กรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำและปานกลาง เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5 พบว่า วิธี Bay-in มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ และเมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1 และ 2 วิธี Bay-in และวิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ แต่เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5 และ 10 พบว่า วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกระดับขนาดตัวอย่าง และเปอร์เซ็นต์การสูญหายของข้อมูล ส่วนในกรณีที่ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกค่าความแปรปรวนของความคลาดเคลื่อน นอกจากนี้ พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า AMSE ของทุกวิธีมีแนวโน้มลดลง และเมื่อความแปรปรวนของความคลาดเคลื่อนมีค่าเพิ่มขึ้น ค่า AMSE ของทุกวิธีเพิ่มขึ้น

5.2 อภิปรายผล

จากการพิจารณาค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีประมาณทั้ง 6 วิธี พบว่ากรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับต่ำและปานกลาง เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 0.5 พบว่า วิธี Bay-in มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่

เนื่องจากวิธีแบบเบส์เมื่อกำหนดการแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ เป็นวิธีที่นำข้อมูลในอดีตมาใช้ประโยชน์ในการประมาณค่า จึงให้การประมาณค่าที่ค่อนข้างแม่นยำ ซึ่งสอดคล้องกับงานวิจัยของ Brandel (2004) พบว่าในสถานการณ์ที่ข้อมูลมีจำนวนจำกัด และมีข้อมูลในอดีตที่เป็นประโยชน์ ควรใช้วิธีแบบเบส์เมื่อกำหนดการแจกแจงก่อนที่ให้สารสนเทศที่เป็นประโยชน์ในการประมาณค่าสัมประสิทธิ์ถดถอย เนื่องจากวิธีแบบเบส์สามารถปรับปรุงการประมาณค่าให้ดีขึ้นเมื่อนำข้อมูลในอดีตใช้ประโยชน์ และเมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 1 และ 2 วิธี Bay-in และวิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ แต่เมื่อความแปรปรวนของความคลาดเคลื่อนเท่ากับ 5 และ 10 พบว่า วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกระดับขนาดตัวอย่างและเปอร์เซ็นต์การสูญหายของข้อมูล ส่วนในกรณีตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันในระดับสูง วิธี BBRI มีประสิทธิภาพดีที่สุดเป็นส่วนใหญ่ในทุกค่าความแปรปรวนของความคลาดเคลื่อน เนื่องจากการนำวิธีเบส์บูตสแตรปมาช่วยในการประมาณค่าพารามิเตอร์จะมีความแม่นยำมาก Clyde M. & Lee H. (2000) ซึ่งสอดคล้องกับงานวิจัยของ Erlangung Z. (2009)

5.3 ข้อเสนอแนะ

1. งานวิจัยนี้ได้ศึกษาวิธีประมาณค่าตัวแปรตามที่สูญหายในการวิเคราะห์ตัวแบบถดถอยเชิงเส้นพหุคูณ โดยใช้จำนวนตัวแปรอิสระเพียง 2 ตัว ดังนั้นผู้ที่สนใจควรศึกษาในกรณีที่มีตัวแปรอิสระมากกว่า 2 ตัว และอาจกำหนดลักษณะข้อมูลให้มีการแจกแจงรูปแบบอื่นๆ

2. ผู้ที่สนใจอาจศึกษาเปรียบเทียบวิธีประมาณค่าสูญหายในกรณีที่มีการสูญหายในตัวแปรอิสระ หรือสูญหายทั้งในตัวแปรตามและตัวแปรอิสระ และควรศึกษาเกณฑ์ที่ใช้ในการเปรียบเทียบเพิ่มเติมจากเกณฑ์ที่ผู้วิจัยใช้ในงานวิจัยนี้

บรรณานุกรม

เกตุจันทร์ จำปาไชยศรี. (2559). *ทฤษฎีสถิต 2 (StatisticalThrory II)*. ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

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

จิรกานต์ นวลละออ. (2552). การเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับตัวแบบพยากรณ์. *วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์ ภาควิชาสถิติประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*, 137 น.

ธรรมรัตน์ กลีบเมฆ, นพคุณ ทองมวล. (2564). *การประมาณค่าสูญหายด้วยวิธีการถดถอยแบบเบย์-บูตสแตรป*. วารสารวิทยาศาสตร์บูรพา มหาวิทยาลัยบูรพา.

นรุตม์ บุตรพลอย. (2553). *การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล*. National Conference on Information Technology. 28, 25-29.

พิมพ์ชนก เชาวณาพรรณ์. (2559). *การเปรียบเทียบการประมาณค่าเฉลี่ยประชากรในการเลือกตัวอย่างสุ่มแบบง่าย กรณีมีข้อมูลสูญหาย*. วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

เรืองลักษณ์ หลำใจซื่อ, อำไพ ทองธีรภาพ และจุฑาภรณ์ สินสมบูรณ์ทอง. (2560). *การเปรียบเทียบวิธีการประมาณค่าสูญหายสาหรับการวิเคราะห์การถดถอยพหุเมื่อตัวแปรตามมีการสูญหายแบบสุ่ม*. วิทยานิพนธ์มหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์.

วราพร ลิ่มชูเชื้อ. (2556). *การเปรียบเทียบวิธีการใส่ค่าสูญหายในตัวแบบการถดถอยเชิงเส้นพหุ เมื่อร้อยละการสูญหายของตัวแปรอิสระและตัวแปรตามต่างกัน สำหรับการสูญหายแบบนอนอิกนอร์เรเบิล*. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์การบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

ศศิธร สมพงศ์นวกิจ. (2555). *การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบร่วม*. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

สุปรียา สระโสม และ ธิดาเดียว มยุรีสวรรค์. (2547). *การเปรียบเทียบวิธีการเติมข้อมูลสูญหายในตัวแปรตามที่เกิดการสูญหายแบบสุ่มสำหรับการถดถอยเชิงเส้นพหุคูณ*. วิทยานิพนธ์มหาบัณฑิต มหาวิทยาลัยขอนแก่น.

เสาวณิต สุขภารังษี. (2546). *การใช้เทคนิคการพยากรณ์ร่วมด้วยตัวถ่วงน้ำหนัก*. วารสารพัฒนาเทคนิคศึกษา, 15(16), 60-65.

อัชฌา อระวีพร, (2555). *การวิเคราะห์เบส์จากโปรแกรมวินบั๊กสู่โปรแกรมอาร์.* วารสาร NU Science Journal, 9(1), 30- 44.

Albert J. (2009). *Bayesian Computation with R.* Springer Dordrecht Heidelberg London New York.

Brandel J. (2004). *Empirical Bayes methods for missing data analysis*. Uppsala University. Research.

Donald B. Rubin. (1981). *The Bayesian Bootstrap*. *The Annals of Statistics*. Institute of Mathematical Statistics, 130-134. from https://www.jstor.org/stable/2240875

Erlangung Z. (2009). *Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. University Bamberg. Research.

Goldstein H., Carpenter J. and Michael G. Kenward, (2018). *Bayesian models for weighted data with missing values: a bootstrap approach*. London School of Hygiene and Tropical Medicine and University College London, UK

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

Clyde M. & Lee H. (2000). *Bagging and the Bayesian Bootstrap.* Retrieved Jan, 2019, form https://www.researchgate.net/publication/2469163.

Sujit K. Ghosh. (2015). *Bayesian Imputation Methods for Missing Data*. Department of Statistics North Carolina State University. from https://www.researchgate.net/publication/2346023