

การเปรียบเทียบประสิทธิภาพตัวแบบการถดถอยเชิงเส้น ภายใต้วิธีบูตสแตรป์แบบไม่อิงพารามิเตอร์

Performance Comparison of the Linear Regression Model Based on Nonparametric Bootstrap Methods

อริสรา อรุณนพรัตน์, อริสสา คำอ่อน, นันทิชา แสนใจธรรม และ อังคณา เกาะแก้ว^{*}

Arisara Arunnobpharat, Arissa Damon, Nanticha Sanjaitham and Angkana Kokaew^{*}

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประเทศไทย

Department of Mathematics, Faculty of Science, Srinakharinwirot University, Thailand

Received : 3 April 2025, Received in revised form : 9 July 2025, Accepted : 14 July 2025

Available online : 24 July 2025

บทคัดย่อ

วัตถุประสงค์และที่มา : การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติที่ใช้กันอย่างแพร่หลายในการทำนายหรือพยากรณ์ค่าของตัวแปรตามจากค่าที่กำหนดให้ของตัวแปรอิสระ โดยการวิเคราะห์การถดถอยเชิงเส้นนั้นมีข้อสมมติเบื้องต้นเกี่ยวกับความคลาดเคลื่อนหลายประการ ซึ่งในบางสถานการณ์ข้อมูลที่นำมาใช้อาจไม่เป็นไปตามข้อสมมติเบื้องต้นทำให้เกิดความคลาดเคลื่อนหรือความผิดพลาดในการพยากรณ์ได้ เพื่อแก้ปัญหาดังกล่าว นักวิจัยจึงได้พัฒนาตัวแบบการถดถอยแบบไม่อิงพารามิเตอร์ขึ้นมา ซึ่งไม่ต้องคำนึงถึงข้อสมมติเบื้องต้นดังกล่าว ทำให้มีความยืดหยุ่นมากขึ้น การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของตัวแบบการถดถอยเชิงเส้นภายใต้วิธีบูตสแตรป์แบบไม่อิงพารามิเตอร์ 2 วิธี ได้แก่ วิธีบูตสแตรป์แบบเปอร์เซ็นไทล์ หรือที่รู้จักกันในชื่อวิธีบูตสแตรป์พี และวิธีบูตสแตรป์แบบสตีเวนสันหรือวิธีบูตสแตรป์ที ซึ่งวิธีบูตสแตรป์พีเป็นเทคนิคที่ได้รับการยอมรับอย่างกว้างขวางเนื่องจากมีรูปแบบของการสุ่มข้อมูลที่ซับซ้อน โดยการสุ่มตัวอย่างข้อมูลซ้ำด้วยการแทนที่คืนเพื่อสร้างตัวอย่างบูตสแตรป์หลาย ๆ ตัวอย่าง จากนั้นทำการคำนวณการแจกแจงของตัวประมาณ วิธีนี้มีประโยชน์อย่างยิ่งเมื่อตัวอย่างมีขนาดเล็กและการแจกแจงของข้อมูลไม่เป็นที่รู้จัก ในทางกลับกัน วิธีบูตสแตรป์ที เป็นการปรับปรุงวิธีบูตสแตรป์พีโดยการนำแนวคิดของการสตีเวนสันมาใช้ ซึ่งเกี่ยวข้องกับการปรับตัวอย่างบูตสแตรป์ตามค่าความคลาดเคลื่อนมาตรฐานของตัวประมาณ ทำให้ได้ช่วงความเชื่อมั่นที่แม่นยำมากขึ้นและลดความเอนเอียง

วิธีดำเนินการวิจัย : งานวิจัยนี้ทำการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โลภายใต้ค่าความคลาดเคลื่อน (ϵ) มีการแจกแจงเบ้ปกติ (Skew-Normal Distributions : SN) 3 ลักษณะ คือ เบ้ซ้าย เบ้ขวา และสมมาตร ซึ่งเป็นการแจกแจงที่พบได้บ่อยในข้อมูลจริง โดยมีพารามิเตอร์ 3 ตัว ได้แก่ พารามิเตอร์บ่งตำแหน่ง (Location Parameter : μ) พารามิเตอร์บ่งขนาด (Scale Parameter : σ^2) และพารามิเตอร์บ่งรูปร่าง (Shape Parameter : λ) โดยพารามิเตอร์สำหรับการจำลองข้อมูลถูกเลือกเพื่อสะท้อนถึงระดับความเบ้และความแปรปรวนที่แตกต่างกัน ทำให้การประเมินประสิทธิภาพในการวิเคราะห์การถดถอยภายใต้วิธีบูตสแตรป์มีความครอบคลุมกับลักษณะการแจกแจงทั้งสามรูปแบบ กำหนดให้ $\mu = 0$, $\sigma^2 = 1, 2, 5, 10$ และ $\lambda = -1, 0, 1$ นั่นคือ $\epsilon \sim \text{SN}(0, \sigma^2, \lambda)$ และจำลองข้อมูลสำหรับตัวแปรอิสระภายใต้การแจกแจงปกติ โดยมีพารามิเตอร์ 2 ตัว ได้แก่ ค่าเฉลี่ย (μ_X) และค่าความแปรปรวน (σ_X^2) กำหนดให้ $\mu_X = 3$ และ $\sigma_X^2 = 2.25$

นั่นคือ $X \sim N(3, 2.25)$ ขนาดตัวอย่างที่ใช้ในจำลอง คือ 20, 30, 50 และ 100 ทำการสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยและสมการถดถอยด้วยวิธีบูตสแตรป์ทั้งสองวิธี ที่ระดับความเชื่อมั่น 95% โดยกระทำซ้ำ 1,000 รอบในแต่ละสถานการณ์ จากนั้นทำการเปรียบเทียบประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอย โดยใช้ค่าสัมประสิทธิ์การถดถอยเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย และค่ารากที่สองของคลาดเคลื่อนกำลังสองเฉลี่ยของสัมประสิทธิ์การถดถอย เป็นค่าเกณฑ์ในการเปรียบเทียบตัวประมาณแบบจุด และประสิทธิภาพของตัวประมาณแบบช่วงถูกเปรียบเทียบโดยใช้ค่าความน่าจะเป็นคัมรวม ค่าความกว้างเฉลี่ยของช่วง และค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วง อีกทั้ง การศึกษาครั้งนี้ได้ประยุกต์ใช้ตัวแบบการถดถอยบูตสแตรป์ทั้งสองวิธีกับข้อมูลจริง 2 ชุด ได้แก่ ชุดข้อมูลน้ำหนักทารกแรกเกิด และชุดข้อมูลเกรดเฉลี่ย โดยเปรียบเทียบประสิทธิภาพของสมการถดถอยที่ดีที่สุดจากเกณฑ์ค่าสัมประสิทธิ์การกำหนด เกณฑ์สารสนเทศของอะกะอิเกะ เกณฑ์สารสนเทศของเบส์ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย

ผลการวิจัย : จากการศึกษาจำลองข้อมูลเชิงสถิติเพื่อเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอย พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นค่าความน่าจะเป็นคัมรวมมีค่าเพิ่มขึ้น ยกเว้นกรณีที่ค่าคลาดเคลื่อนมีการแจกแจงแบบเบ้สำหรับพารามิเตอร์ค่าคงที่ ค่าความกว้างเฉลี่ยของช่วง และค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงมีแนวโน้มลดลงในทุกชุดพารามิเตอร์ โดยส่วนมากวิธีบูตสแตรป์ที่มีค่าความน่าจะเป็นคัมรวมมากที่สุด และผลการเปรียบเทียบประสิทธิภาพของสมการถดถอยบูตสแตรป์ในการประมาณค่าพารามิเตอร์ พบว่าในทุกขนาดตัวอย่างค่าสัมประสิทธิ์การถดถอยเฉลี่ยซึ่งเป็นค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีบูตสแตรป์มีค่าใกล้เคียงค่าพารามิเตอร์ และเมื่อขนาดตัวอย่างเพิ่มขึ้นค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย และค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยลดลงในทุกกรณี นั่นคือสมการถดถอยจะมีประสิทธิภาพเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น นอกจากนี้ในทุกขนาดตัวอย่างสมการถดถอยบูตสแตรป์ที่ให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าสมการถดถอยบูตสแตรป์ที่ ซึ่งบ่งชี้ถึงประสิทธิภาพที่ดีกว่า ยกเว้นกรณีที่ค่าความคลาดเคลื่อนสุ่มมีการแจกแจงแบบสมมาตรและขนาดตัวอย่างเล็ก จะพบว่าวิธีบูตสแตรป์ที่มีประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอยสูงกว่า

สรุปผลการวิจัย : การเปรียบเทียบสมการถดถอยบูตสแตรป์เผยให้เห็นว่าวิธีทั้งสองให้ค่าประมาณที่ใกล้เคียงกับพารามิเตอร์จริง ซึ่งบ่งชี้ถึงประสิทธิภาพในการจัดการกับการแจกแจงความคลาดเคลื่อนที่แตกต่างกัน อย่างไรก็ตามวิธีบูตสแตรป์ที่แสดงให้เห็นถึงประสิทธิภาพที่เหนือกว่าวิธีบูตสแตรป์ฟอยอย่างต่อเนื่อง โดยเฉพาะในแง่ของค่าความน่าจะเป็นคัมรวมและค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ซึ่งเป็นมาตรวัดสำคัญของความแม่นยำในการประมาณค่าสัมประสิทธิ์การถดถอย บ่งชี้ว่าวิธีบูตสแตรป์ที่มีความน่าเชื่อถือมากกว่า ทำให้เป็นทางเลือกที่น่าเชื่อถือในการสร้างช่วงความเชื่อมั่นและการประมาณค่าสัมประสิทธิ์การถดถอย และเมื่อขนาดตัวอย่างเพิ่มขึ้นความแม่นยำและความน่าเชื่อถือของวิธีบูตสแตรป์จะดีขึ้น ทำให้ช่วงความเชื่อมั่นแคบลงและการประมาณค่าสัมประสิทธิ์การถดถอยแม่นยำมากขึ้น นอกจากนี้ ผลลัพธ์ที่ได้จากการประยุกต์ใช้กับข้อมูลจริงนั้นสอดคล้องกับการจำลองข้อมูลเชิงสถิติภายใต้การแจกแจงแบบเบ้ปกติ ดังนั้น วิธีบูตสแตรป์จึงเป็นวิธีทางเลือกในการประมาณค่าช่วงความเชื่อมั่นและการประมาณค่าสัมประสิทธิ์การถดถอยที่ทำให้สมการถดถอยที่ได้นั้นมีประสิทธิภาพในการพยากรณ์ โดยเฉพาะเมื่อข้อสมมติของวิธีอิงพารามิเตอร์ไม่เป็นจริง ผลการศึกษานี้มีความสำคัญและประโยชน์ต่อทั้งนักวิจัยและผู้ปฏิบัติงาน ซึ่งเน้นถึงการพิจารณาขนาดตัวอย่างและการแจกแจงความคลาดเคลื่อนอย่างรอบคอบเพื่อเลือกวิธีการถดถอยที่เหมาะสม

คำสำคัญ : วิธีบูตสแตรป์ ; ช่วงความเชื่อมั่น ; เทคนิคมอนติคาร์โล ; การวิเคราะห์การถดถอย ; การแจกแจงเบ้ปกติ

Abstract

Background and Objectives : Regression analysis is a widely used statistical method for predicting or forecasting the value of a dependent variable from given values of independent variables. Linear regression analysis has several underlying assumptions about errors, but in some situations, the data may not meet these assumptions, leading to prediction errors. To address this issue, researchers have developed nonparametric regression models that do not rely on assumptions about the distribution of errors, providing greater flexibility. The objective of this research was to investigate and compare the performance of linear regression models under nonparametric bootstrap methods. This research utilized two different nonparametric bootstrap methods; the percentile bootstrap method (also known as the bootstrap-p method) and the studentized bootstrap method (or bootstrap-t method). The bootstrap-p method is widely accepted due to its simple resampling technique, which involves repeatedly sampling with replacement to create multiple bootstrap samples and then calculating the distribution of the estimator. This method is particularly useful when the sample size is small, and the data distribution is unknown. In contrast, the bootstrap-t method enhances the bootstrap-p method by incorporating the concept of studentization. This involves adjusting the bootstrap samples according to the standard error of the estimator, resulting in more accurate confidence intervals and reduced bias.

Methodology : The performance of the bootstrap regression models was examined using Monte Carlo simulation, based on the distribution of error terms (ϵ) with three types of skew-normal distributions; left-skewed, right-skewed, and symmetric, which are commonly observed in real data. The parameters of skew-normal distribution include a location parameter (μ), a scale parameter (σ^2), and a shape parameter (λ). These parameters for data simulation were selected to reflect different levels of skewness and variance, ensuring a comprehensive evaluation of the performance of bootstrap regression analysis across the three distribution types. The parameters were set as follows: $\mu = 0$, $\sigma^2 = 1, 2, 5, 10$, and $\lambda = -1, 0, 1$, i.e., $\epsilon \sim \text{SN}(0, \sigma^2, \lambda)$. For the independent variable, data was simulated under normal distribution with parameters $\mu_X = 3$ and $\sigma_X^2 = 2.25$, i.e., $X \sim \text{N}(3, 2.25)$. Sample sizes of 20, 30, 50, and 100 were used. Confidence intervals for regression coefficients and regression models were constructed using both bootstrap methods at a 95% confidence level, with 1,000 repetitions for each situation. The performance of the regression coefficient estimates was compared using the coefficient mean, the standard deviation of the regression coefficient, and the root mean squared error of the regression coefficient were presented as the criteria of point estimation. Also, the performance of interval estimation was compared with coverage probability, average width, and standard deviation of interval width. Additionally, both bootstrap regression models were applied to two real datasets: newborn birth weights and grade point averages, comparing the performance of the best regression models based on the coefficient of determination, Akaike's information criterion, Bayesian information criterion, and mean squared error.

Main Results : The simulation study aimed at comparing the efficiency of confidence intervals for regression coefficients found that, as the sample size increases, the overall coverage probability increases, except in cases where the error distribution was skewed for the intercept parameter. Additionally, the average width of the interval, and the standard deviation of the interval width tend to decrease in all parameter sets. In most cases, the bootstrap-t method consistently achieved the highest coverage probability. The comparison of the performance of bootstrap regression models in parameter estimation indicated that, for all sample sizes, the mean regression coefficient, which is the bootstrap estimate of the regression coefficient, closely approximated the parameter values. Additionally, as the sample size increased, the standard deviation of the regression coefficient and the root mean squared error of the regression coefficient decreased in all cases. This implies that the efficiency of the regression models improves with larger sample sizes. Furthermore, for all sample sizes, the bootstrap-t regression models provided a lower root mean squared error of the regression coefficient compared to the bootstrap-p regression models, indicating better performance. However, in cases where the random error terms followed a symmetric distribution and the sample size was small, the bootstrap-p method demonstrated higher efficiency in estimating the regression coefficients.

Conclusions : The comparison of bootstrap regression models revealed that both methods provided estimates close to the true parameters, indicating their effectiveness in handling different error distributions. However, the bootstrap-t method consistently demonstrated superior performance, particularly in terms of the coverage probability and the root mean squared error, which are key measures of accuracy in estimating regression coefficients. This suggests that the bootstrap-t method is more reliable, making it a more credible choice for constructing confidence intervals and estimating regression coefficients. As the sample size increased, the accuracy and reliability of the bootstrap methods improved, resulting in narrower confidence intervals and more precise regression coefficient estimates. Additionally, the results obtained from applying the bootstrap methods to real data were consistent with the simulations study under skew-normal distributions. Therefore, bootstrap methods are alternative methods for estimating confidence intervals and regression coefficients. They enhance the predictive efficiency of regression models, especially when parametric assumptions are violated. This study is significant and beneficial for both researchers and practitioners. It emphasizes the thorough consideration of sample size and error distribution to select the appropriate regression method.

Keywords: Bootstrap method; confidence interval ; Monte Carlo simulation ; regression model ; skew-normal distribution

*Corresponding author. E-mail: angkanak@g.swu.ac.th

Introduction

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรเพื่อสร้างสมการพยากรณ์ ซึ่งใช้กันอย่างแพร่หลายในการทำนายและการพยากรณ์ค่าของตัวแปรตามจากค่าที่กำหนดให้ของตัวแปรอิสระ โดยสมการถดถอยเชิงเส้นมีข้อสมมติเบื้องต้น ได้แก่ ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นตรง ค่าเฉลี่ยของความคลาดเคลื่อนเท่ากับศูนย์ ความแปรปรวนของความคลาดเคลื่อนคงที่ ความคลาดเคลื่อนมีการแจกแจงปกติ และตัวแปรอิสระไม่มีความสัมพันธ์ซึ่งกันและกัน (Multicollinearity) ซึ่งบางสถานการณ์ข้อมูลที่น่าสนใจอาจไม่เป็นไปตามข้อสมมติเบื้องต้นทำให้เกิดความคลาดเคลื่อนในการพยากรณ์ได้ อีกทั้งตัวแบบที่มีประสิทธิภาพนั้นมักจะเกิดจากการวิเคราะห์ภายใต้ข้อมูลที่มีขนาดใหญ่ แต่ในบางปัจจัยที่ทำให้ไม่สามารถเก็บข้อมูลขนาดใหญ่ได้ เช่น ข้อมูลที่มีข้อจำกัดของระยะเวลา ค่าใช้จ่าย และความยากในการเข้าถึงข้อมูล เพื่อแก้ปัญหการวิเคราะห์การถดถอยที่ไม่ตรงตามข้อสมมติเบื้องต้น จึงได้มีการสร้างตัวแบบการถดถอยแบบไม่อิงพารามิเตอร์ (Nonparametric Regression Model) ขึ้น

วิธีบูตสแตรป์ (Bootstrap Method) ถูกเผยแพร่ในปี ค.ศ. 1979 โดย Bradley Efron (Efron, 1979) เป็นวิธีที่นิยมนำมาใช้ในการประมาณค่าพารามิเตอร์เมื่อข้อมูลตัวอย่างมีขนาดเล็กหรือไม่ทราบการแจกแจงของข้อมูลประชากร ซึ่งวิธีบูตสแตรป์เป็นเทคนิคการสุ่มตัวอย่างซ้ำแบบแทนที่เพื่อสร้างการแจกแจงของตัวสถิติโดยไม่ต้องคำนึงถึงขนาดตัวอย่างและข้อตกลงเบื้องต้นของการแจกแจงข้อมูล ทำให้สามารถหาค่าประมาณพารามิเตอร์ ค่าประมาณของความคลาดเคลื่อนมาตรฐาน และช่วงความเชื่อมั่นของพารามิเตอร์ที่สนใจได้ อีกทั้งยังสามารถนำมาประยุกต์ใช้ในการวิเคราะห์การถดถอยได้อีกด้วย เรียกว่า การวิเคราะห์การถดถอยบูตสแตรป์ (Bootstrap Regression Analysis)

ในงานวิจัยนี้ ผู้วิจัยทำการศึกษาวิธีบูตสแตรป์แบบไม่อิงพารามิเตอร์สำหรับวิเคราะห์การถดถอยเชิงเส้น 2 วิธี ได้แก่ วิธีบูตสแตรป์แบบเปอร์เซ็นต์ไทล์ (Percentile Bootstrap Method) หรือเรียกว่า วิธีบูตสแตรป์พี (Bootstrap-p Method) (Efron, 1982; Hall, 1988) ซึ่งเป็นวิธีที่นิยมใช้กันอย่างกว้างขวาง เนื่องจากมีรูปแบบของการสุ่มข้อมูลที่ไม่ซับซ้อน และวิธีบูตสแตรป์แบบสตีวเดนซ์ (Studentized Bootstrap Method) หรือวิธีบูตสแตรป์ที (Bootstrap-t Method) (Efron & Tibshirani, 1994; Wilcox, 2017) เป็นวิธีที่พัฒนามาจากวิธีบูตสแตรป์พีซึ่งมีประสิทธิภาพมากกว่า โดยนำหลักการสุ่มซ้ำแบบบูตสแตรป์มาประยุกต์ใช้เพื่อประมาณค่าพารามิเตอร์แบบช่วง ดังนั้นวัตถุประสงค์หลักของงานวิจัยนี้ คือ เพื่อเปรียบเทียบประสิทธิภาพของสมการถดถอยภายใต้วิธีบูตสแตรป์ 2 วิธี คือ วิธีบูตสแตรป์พี และวิธีบูตสแตรป์ที โดยทำการศึกษาการจำลองเชิงสถิติด้วยเทคนิคมอนติคาร์โล (Monte Carlo Simulation) เมื่อค่าความคลาดเคลื่อนมีการแจกแจงเบ้ปกติ (Skew-Normal Distribution : SN) 3 ลักษณะ ได้แก่ เบ้ซ้าย เบ้ขวา และสมมาตร (Figueiredo & Gomes, 2012; Thiutad, 2021) เพื่อวัดประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอยทั้งการประมาณค่าแบบจุดและการประมาณค่าแบบช่วง ด้วยค่าเกณฑ์สัมประสิทธิ์การถดถอยเฉลี่ย (Mean Coefficient) ค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย (Standard Deviation of Regression Coefficient) และค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error) ของสัมประสิทธิ์การถดถอย ซึ่งเป็นเกณฑ์ที่ใช้ในการเปรียบเทียบตัวประมาณแบบจุด และพิจารณาค่าความน่าจะเป็นคัมรวม (Coverage Probability) (Kokaew *et al.*, 2021) ค่าความกว้างเฉลี่ยของช่วง (Average Width) และค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วง (Standard Deviation of Interval Width) ซึ่งเป็นเกณฑ์ที่ใช้ในการเปรียบเทียบตัวประมาณแบบช่วง (Kashif *et al.*, 2017) นอกจากนี้

นำวิธีการถดถอยที่ศึกษาไปประยุกต์ใช้กับข้อมูลจริงและเปรียบเทียบประสิทธิภาพของสมการถดถอยด้วยค่าสัมประสิทธิ์การกำหนด (Coefficient of Determination) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error) เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's Information Criterion) และเกณฑ์สารสนเทศของเบส์ (Bayesian Information Criterion) พร้อมทั้งประยุกต์กับข้อมูลจริง

Methodology

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของวิธีบูตสแตรป์แบบไม่อิงพารามิเตอร์ 2 วิธี สำหรับการสร้างสมการถดถอยเชิงเส้น ได้แก่ วิธีบูตสแตรป์พี และวิธีบูตสแตรป์ที ซึ่งวิธีการดำเนินงานวิจัยในครั้งนี้เป็นการจำลองข้อมูลโดยใช้เทคนิคมอนติคาร์โล และนำตัวแบบการถดถอยด้วยวิธีบูตสแตรป์มาประยุกต์ใช้กับชุดข้อมูลจริง โดยใช้โปรแกรม R (R Core Team, 2022; Dikta & Scheer, 2021) มีขั้นตอนดังต่อไปนี้

(1) กำหนดตัวแบบการถดถอยเชิงเส้นอย่างง่าย $Y = \beta_0 + \beta_1 X + \epsilon$

(2) จำลองข้อมูลขนาด n ตัว คือ 20, 30, 50 และ 100 กำหนดให้ $\beta_0 = 5$ และ $\beta_1 = 1$ นั่นคือเส้นตรงที่มีความชันเท่ากับ 1 และจุดตัดแกน y ที่ 5 เป็นกรณีพื้นฐานที่ง่ายต่อการตีความ

(3) จำลองค่าความคลาดเคลื่อน (ϵ) ภายใต้การแจกแจงเบ้ปกติ (Eck, 2018) โดยมีพารามิเตอร์คือ พารามิเตอร์บ่งตำแหน่ง (Location parameter : μ) พารามิเตอร์บ่งขนาด (Scale parameter : σ^2) และพารามิเตอร์บ่งรูปร่าง (Shape parameter : λ) โดยกำหนดให้ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์ $\mu = 0$ ความแปรปรวนของความคลาดเคลื่อนมีระดับต่างกันครอบคลุมทั้งค่าสูงและต่ำ ได้แก่ $\sigma^2 = 1, 2, 5, 10$ และกำหนดลักษณะรูปร่างของการแจกแจงของความคลาดเคลื่อนให้หลากหลาย ได้แก่ เบ้ซ้าย สมมาตร และเบ้ขวา ดังนี้ $\lambda = -1, 0, 1$ นั่นคือ $\epsilon \sim SN(0, \sigma^2, \lambda)$ และจำลองข้อมูลของตัวแปรอิสระภายใต้การแจกแจงปกติ โดยมีพารามิเตอร์ 2 ตัว คือ ค่าเฉลี่ย (μ_x) และค่าความแปรปรวน (σ_x^2) เพื่อไม่ให้ตัวแปรอิสระทับซ้อนกับค่าคงที่มากเกินไป ทำให้การวิเคราะห์ผลของ β_1 ชัดเจนขึ้น กำหนดให้ $\mu_x = 3$ และเพื่อให้เห็นผลของการเปลี่ยนแปลงในตัวแปรอิสระต่อค่าตอบสนองได้ชัดเจน โดยไม่ทำให้ข้อมูลกระจายมากจนยากต่อการวิเคราะห์ กำหนดให้ $\sigma_x^2 = 2.25$ นั่นคือ $X \sim N(3, 2.25)$

(4) สร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยและสมการถดถอยด้วยวิธีต่อไปนี้

4.1 วิธีบูตสแตรป์พีแบบไม่อิงพารามิเตอร์ (Nonparametric Bootstrap-p Method)

4.2 วิธีบูตสแตรป์ทีแบบไม่อิงพารามิเตอร์ (Nonparametric Bootstrap-t Method)

(5) ระดับความเชื่อมั่นที่กำหนดเท่ากับ 95 %

(6) วัดประสิทธิภาพของการประมาณช่วงความเชื่อมั่นด้วยค่าความน่าจะเป็นคัมรวม ค่าความกว้างเฉลี่ยของช่วง และค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วง พร้อมทั้งเปรียบเทียบประสิทธิภาพของสมการถดถอย โดยพิจารณาจากค่าสัมประสิทธิ์การถดถอยเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย และค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของสัมประสิทธิ์การถดถอย

(7) จำนวนรอบของบูตสแตรป์ที่ให้ความแม่นยำและไม่ใช้เวลาคำนวณมากเกินไป คือ $B_1 = 1,000$ และ $B_2 = 50$

(8) จำลองด้วยเทคนิคมอนติคาร์โลโดยกระทำซ้ำ $R = 1,000$ รอบ ในแต่ละสถานการณ์

(9) ประยุกต์ใช้ตัวแบบการถดถอยบูตสแตรป์ทั้งสองวิธีกับข้อมูลจริง 2 ชุด ได้แก่ ชุดข้อมูลน้ำหนักทารกแรกเกิด และชุดข้อมูลเกรดเฉลี่ย โดยเปรียบเทียบประสิทธิภาพของสมการถดถอยที่ดีที่สุดจากเกณฑ์ค่าสัมประสิทธิ์การกำหนด เกณฑ์สารสนเทศของอะกะอิเกะ เกณฑ์สารสนเทศของเบส์ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการวิเคราะห์การถดถอยด้วยวิธีบูตสแตรป์ 2 วิธี มีขั้นตอนการคำนวณดังนี้

(1) วิธีบูตสแตรป์ (Bootstrap-p Method) (Davison & Kuonen, 2002)

การวิเคราะห์การถดถอยบูตสแตรป์มีขั้นตอนการวิเคราะห์ ดังนี้
ขั้นตอนที่ 1 : จำลองข้อมูลตัวแปรอิสระ (X_i) และค่าความคลาดเคลื่อน (ε_i) ขนาด n ค่าสังเกต จากนั้นหาค่าสังเกตของ ตัวแปรอิสระ (Y_i) จากสมการ

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

โดยกำหนดค่าพารามิเตอร์ β_0 และ β_1 เมื่อ $i = 1, 2, \dots, n$

ขั้นตอนที่ 2 : จากค่าสังเกต (X_i) และค่าสังเกต (Y_i) นำมาสร้างสมการถดถอย จะได้สมการดังนี้

$$\hat{Y}_i = b_0 + b_1 X_i \quad (2)$$

โดยที่ $\hat{Y}_i$ คือ ค่าพยากรณ์ ลำดับที่ i เมื่อ $i = 1, 2, \dots, n$

b_0 และ b_1 คือค่าประมาณของพารามิเตอร์ β_0 และ β_1 ตามลำดับ

ขั้นตอนที่ 3 : จากสมการถดถอยขั้นตอนที่ 2 หาค่าพยากรณ์ ($\hat{Y}_i$) เมื่อ $i = 1, 2, \dots, n$ โดยการแทนค่า X_i ลงในสมการ (2)

ขั้นตอนที่ 4 : นำค่าสังเกต (Y_i) และค่าพยากรณ์ ($\hat{Y}_i$) มาคำนวณค่าส่วนเหลือ (e_i) จากสูตร $e_i = Y_i - \hat{Y}_i$

ขั้นตอนที่ 5 : กำหนดรอบบูตสแตรป์เท่ากับ B_1 และทำการสุ่มค่าส่วนเหลือ (e_i) โดยตัวอย่างชุดที่ j จะทำการสุ่มทีละ 1 ค่าแบบแทนที่คืน จำนวน n ครั้ง จะได้ค่าส่วนเหลือบูตสแตรป์ชุดที่ j คือ $e_i^{(j)}$ เมื่อ $i = 1, 2, \dots, n$ และ $j = 1, 2, \dots, B_1$

ขั้นตอนที่ 6 : คำนวณค่า Y_i ของบูตสแตรป์ (Y_i^*) จากสมการ

$$Y_i^* = \hat{Y}_i + e_i^* \quad (3)$$

ขั้นตอนที่ 7 : นำค่า Y_i^* และ X_i ขนาด n ประมาณค่าสัมประสิทธิ์การถดถอย จากสูตร

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}^* \quad (4)$$

จะได้ค่าประมาณสัมประสิทธิ์การถดถอยของบูตสแตรป์ $b_c^{*(i)}$ เมื่อ $c = 0, 1$

ขั้นตอนที่ 8 : ทำซ้ำขั้นตอนที่ 5-7 จำนวน B_1 รอบ จะได้ค่าประมาณสัมประสิทธิ์การถดถอย $b_c^{*(1)}, b_c^{*(2)}, \dots, b_c^{*(B_1)}$

ขั้นตอนที่ 9 : นำค่า $b_c^{*(1)}, b_c^{*(2)}, \dots, b_c^{*(B_1)}$ มาคำนวณช่วงความเชื่อมั่น โดยช่วงความเชื่อมั่น $(1 - \alpha)100\%$ ของ β_c มีรูปแบบดังนี้

$$(b_c^{*(\alpha/2)}, b_c^{*(1-\alpha/2)}) \quad (5)$$

โดยที่ $b_c^{*(\alpha/2)}$ คือ ค่าควอนไทล์ที่ $(\alpha/2)$ ของค่าประมาณ $b_c^{*(j)}$
และ $b_c^{*(1-\alpha/2)}$ คือ ค่าควอนไทล์ที่ $(1 - \alpha/2)$ ของค่าประมาณ $b_c^{*(j)}$ เมื่อ $j = 1, 2, \dots, B_1$ และ $c = 0, 1$
ขั้นตอนที่ 10 : นำค่า $b_c^{*(1)}, b_c^{*(2)}, \dots, b_c^{*(B_1)}$ มาสร้างสมการถดถอยโดยประมาณค่าสัมประสิทธิ์การถดถอยจาก

$$\bar{b}_c^* = \frac{\sum_{j=1}^{B_1} b_c^{*(j)}}{B_1} \quad (6)$$

โดยที่ $b_c^{*(j)}$ คือ ค่าประมาณสัมประสิทธิ์การถดถอยของบูตสแตรป์รอบที่ j เมื่อ $j = 1, 2, \dots, B_1$ และ $c = 0, 1$
จะได้สมการถดถอยของบูตสแตรป์ ดังนี้

$$\hat{Y}_i^* = \bar{b}_0^* + \bar{b}_1^* X_i \quad (7)$$

(2) วิธีบูตสแตรป์ที (Bootstrap-t Method)

การวิเคราะห์การถดถอยบูตสแตรป์ที่มีขั้นตอนการวิเคราะห์ ดังนี้

ขั้นตอนที่ 1 : จำลองข้อมูลตัวแปรอิสระ (X_i) และค่าความคลาดเคลื่อน (ε_i) ขนาด n ค่าสังเกต จากนั้นหาค่าสังเกต (Y_i) จากสมการ

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (8)$$

โดยกำหนดค่าพารามิเตอร์ β_0 และ β_1 เมื่อ $i = 1, 2, \dots, n$

ขั้นตอนที่ 2 : จากค่าสังเกต (X_i) และค่าสังเกต (Y_i) นำมาสร้างสมการถดถอย จะได้สมการดังนี้

$$\hat{Y}_i = b_0 + b_1 X_i \quad (9)$$

โดยที่ $\hat{Y}_i$ คือ ค่าพยากรณ์ ลำดับที่ i เมื่อ $i = 1, 2, \dots, n$

b_0 และ b_1 คือค่าประมาณของพารามิเตอร์ β_0 และ β_1 ตามลำดับ

ขั้นตอนที่ 3 : จากสมการถดถอยขั้นตอนที่ 2 หาค่าพยากรณ์ ($\hat{Y}_i$) เมื่อ $i = 1, 2, \dots, n$ โดยการแทนค่า X_i ลงในสมการ (9)

ขั้นตอนที่ 4 : นำค่าสังเกต (Y_i) และค่าพยากรณ์ ($\hat{Y}_i$) มาคำนวณค่าส่วนเหลือ (e_i) จากสูตร $e_i = Y_i - \hat{Y}_i$

ขั้นตอนที่ 5 : กำหนดรอบบูตสแตรป์เท่ากับ B_1 และทำการสุ่มค่าส่วนเหลือ (e_i) โดยตัวอย่างชุดที่ j จะทำการสุ่มทีละ 1

ค่าแบบแทนที่คืน จำนวน n ครั้ง จะได้ค่าส่วนเหลือของบูตสแตรป์ชุดที่ j คือ $e_i^{*(j)}$ เมื่อ $i = 1, 2, \dots, n$ และ $j = 1, 2, \dots, B_1$

ขั้นตอนที่ 6 : คำนวณค่า Y_i^* ของบูตสแตรป์ (Y_i^*) จากสมการ

$$Y_i^* = \hat{Y}_i + e_i^* \quad (10)$$

ขั้นตอนที่ 7 : นำค่า Y_i^* และ X_i ขนาด n ประมาณค่าสัมประสิทธิ์การถดถอย จากสูตร

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}^* \quad (11)$$

จะได้ค่าประมาณสัมประสิทธิ์การถดถอย $b_c^{*(j)}$ เมื่อ $c = 0, 1$

ขั้นตอนที่ 8 : กำหนดรอบบูตสแตรป์เท่ากับ B_2 ทำการสุ่มค่าส่วนเหลือจาก $e_1^{*(j)}, e_2^{*(j)}, \dots, e_n^{*(j)}$ โดยชุดที่ m จะทำการสุ่มตัวอย่างทีละ 1 ค่าแบบแทนที่คืนจำนวน n ครั้ง จะได้ค่าส่วนเหลือของบูตสแตรป์ชุดที่ m คือ $e_i^{** (m)}$ เมื่อ $i = 1, 2, \dots, n$

ขั้นตอนที่ 9 : คำนวณค่า Y_i ของบูตสแตรป์ (Y_i^{**}) จากสมการ $Y_i^{**} = \hat{Y}_i + e_i^{**}$

ขั้นตอนที่ 10 : นำค่า Y_i^{**} และ X_i ขนาด n ประมาณค่าสัมประสิทธิ์การถดถอย จากสูตร

$$b^{**} = (X'X)^{-1} X'Y^{**} \quad (12)$$

จะได้ค่าประมาณสัมประสิทธิ์การถดถอย $b_c^{** (m)}$ เมื่อ $c = 0, 1$

ขั้นตอนที่ 11 : ทำซ้ำขั้นตอนที่ 8-10 จำนวน B_2 ครั้ง จะได้ค่าประมาณสัมประสิทธิ์การถดถอย $b_c^{** (m)}$ เมื่อ $m = 1, 2, \dots, B_2$ และ $c = 0, 1$

ขั้นตอนที่ 12 : คำนวณหาค่าคลาดเคลื่อนมาตรฐานของ b_c^* จากสูตร

$$se(b_c^*) = \sqrt{\frac{1}{B_2-1} \sum_{m=1}^{B_2} (b_c^{** (m)} - \bar{b}_c^{**})^2} \text{ เมื่อ } \bar{b}_c^{**} = \sum_{m=1}^{B_2} \frac{b_c^{** (m)}}{B_2} \quad (13)$$

ขั้นตอนที่ 13 : คำนวณ $T_{b_c}^* = \frac{b_c^* - b_c}{se(b_c^*)}$

ขั้นตอนที่ 14 : ทำซ้ำขั้นตอนที่ 5-13 จำนวน B_1 ครั้ง จะได้ค่าประมาณสัมประสิทธิ์การถดถอย $b_c^{*(j)}$

เมื่อ $j = 1, 2, \dots, B_1$ และ $c = 0, 1$

ขั้นตอนที่ 15 : คำนวณหาค่าคลาดเคลื่อนมาตรฐานของ b_c จากสูตร

$$se(b_c) = \sqrt{\frac{1}{B_1-1} \sum_{j=1}^{B_1} (b_c^{*(j)} - \bar{b}_c^*)^2} \text{ เมื่อ } \bar{b}_c^* = \sum_{j=1}^{B_1} \frac{b_c^{*(j)}}{B_1} \quad (14)$$

ขั้นตอนที่ 16 : คำนวณหาช่วงความเชื่อมั่น $(1 - \alpha)100\%$ ของ β_c ดังนี้

$$(b_c + T_{b_c}^{*(\alpha/2)} \cdot se(b_c), b_c + T_{b_c}^{*(1-\alpha/2)} \cdot se(b_c)) \quad (15)$$

โดยที่ $T_{b_c}^{*(\alpha/2)}$ คือ ค่าควอนไทล์ที่ $(\alpha/2)$ ของ $T_{b_c}^{*(j)}$

และ $T_{b_c}^{*(1-\alpha/2)}$ คือ ค่าควอนไทล์ที่ $(1 - \alpha/2)$ ของ $T_{b_c}^{*(j)}$ เมื่อ $j = 1, 2, \dots, B_1$ และ $c = 0, 1$

ขั้นตอนที่ 17 : นำค่า $b_c^{*(1)}, b_c^{*(2)}, \dots, b_c^{*(B_1)}$ มาคำนวณหาสัมประสิทธิ์การถดถอย จากสูตร

$$\bar{b}_c^* = \frac{\sum_{j=1}^{B_1} b_c^{*(j)}}{B_1} \quad (16)$$

โดยที่ $\bar{b}_c^*$ คือ ค่าเฉลี่ยของ b_c^* เมื่อ $c = 0, 1$

และ $b_c^{*(j)}$ คือ สัมประสิทธิ์การถดถอยของบูตสแตรป์รอบที่ $j = 1, 2, \dots, B_1$ และ $c = 0, 1$

จะได้สมการถดถอยของบูตสแตรป์ ดังนี้

$$Y_i^* = \bar{b}_0 + \bar{b}_1 X_i \quad (17)$$

เกณฑ์ที่ใช้ในการศึกษาประกอบด้วย 2 ส่วน ดังนี้

(1) เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น

1.1 ค่าความน่าจะเป็นคัมรวม (Coverage Probability : CP) คือ ความน่าจะเป็นที่ช่วงความเชื่อมั่นหรือขอบเขตความเชื่อมั่นครอบคลุมค่าพารามิเตอร์โดยพิจารณาจากข้อมูลการทำซ้ำหลาย ๆ รอบ (Kokaew *et al.*, 2021) โดยคำนวณจากสูตร

$$CP = \frac{\sum_{i=1}^R I(i, \theta)}{R} \quad (18)$$

โดยที่ $I(i, \theta)$ คือ ค่าที่แสดงถึงการครอบคลุมค่าพารามิเตอร์ของช่วงความเชื่อมั่นที่ใช้ประมาณค่าพารามิเตอร์นั้น ซึ่งจะมีค่าเท่ากับ 1 เมื่อช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์และมีค่าเท่ากับ 0 เมื่อช่วงความเชื่อมั่นไม่ครอบคลุมค่าพารามิเตอร์ เมื่อ $i = 1, 2, \dots, R$ และ R คือ จำนวนรอบการทำซ้ำ

ดังนั้น ถ้าความน่าจะเป็นคัมรวมที่ได้มีค่าใกล้เคียงหรือมากกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด แสดงว่าช่วงความเชื่อมั่นนั้นยังมีประสิทธิภาพ

1.2 ค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วง (Standard Deviation of Interval Width : SD_w) (Kashif *et al.*, 2017) คือ ค่าที่ใช้วัดการกระจายของข้อมูล ดังนั้น ถ้าค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงมีค่าน้อย แสดงว่าข้อมูลส่วนใหญ่อยู่ใกล้ค่าเฉลี่ยมาก ค่าส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงสามารถคำนวณได้ดังนี้

$$SD_w = \sqrt{\frac{1}{R-1} \sum_{i=1}^R (W_i - AW)^2} \quad (19)$$

โดยที่ $W_i = U_i - L_i$ คือ ผลต่างระหว่างค่าขอบเขตบนและค่าขอบเขตล่างของช่วงความเชื่อมั่น เมื่อ $i = 1, 2, \dots, R$

$AW = \frac{\sum_{i=1}^R (U_i - L_i)}{R}$ คือ ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น (Average Width)

การพิจารณาประสิทธิภาพของการประมาณช่วงความเชื่อมั่นของพารามิเตอร์พิจารณาจากวิธีที่ให้ค่า CP สูงสุด ค่า AW และค่า SD_w ที่ให้ค่าน้อยที่สุด จะถือว่าเป็นวิธีการประมาณค่าที่มีประสิทธิภาพ แต่ในบางวิธีที่ให้ค่า AW น้อยก็ให้ค่า CP ที่น้อยด้วย ซึ่งเราจะให้ความสำคัญในการพิจารณาประสิทธิภาพของการประมาณช่วงความเชื่อมั่นจากค่า CP เป็นอันดับแรก (Kokaew *et al.*, 2021) เนื่องจากเป็นค่าที่บ่งบอกถึงสัดส่วนของช่วงความเชื่อมั่นที่ครอบคลุมค่าพารามิเตอร์ แต่ถ้าช่วงความเชื่อมั่นแต่ละวิธีมีค่า CP เท่ากันหรือใกล้เคียงกันมาก จะพิจารณา SD_w และ AW ถัดไป ซึ่งในงานวิจัยนี้กำหนดสัมประสิทธิ์ความเชื่อมั่น 0.95 ดังนั้นค่า CP ที่เหมาะสมควรมีค่า $CP \geq 0.95 - 1.645 \sqrt{\frac{0.95(1-0.95)}{1000}}$ โดยที่ 0.95 คือสัมประสิทธิ์ความเชื่อมั่นที่กำหนด นั่นคือค่า $CP \geq 0.939$ และเมื่อค่า CP มีค่ามาก จะทำให้ช่วงความเชื่อมั่นมีประสิทธิภาพดีขึ้นด้วย

(2) เกณฑ์ในการเปรียบเทียบประสิทธิภาพของตัวแบบถดถอย

2.1 ค่าสัมประสิทธิ์การถดถอยเฉลี่ย (Mean Coefficient : $\bar{b}_c$) เป็นการหาค่ากลางของข้อมูล เพื่อเป็นตัวแทนของข้อมูลทั้งหมด ซึ่งจะใช้ในการนำข้อมูลตัวอย่างแต่ละชุดมาเปรียบเทียบกัน โดยที่ไม่จำเป็นต้องพิจารณาข้อมูลทั้งหมดของแต่ละชุด

$$\bar{b}_c = \frac{1}{R} \sum_{i=1}^R \bar{b}_{c(i)} \quad (20)$$

โดยที่ $\bar{b}_{c(i)}$ คือ ค่าเฉลี่ยของ b_c^* รอบการทำซ้ำที่ i เมื่อ $c = 0, 1$ และ $i = 1, 2, \dots, R$

2.2 ค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย (Standard Deviation of Coefficient Regression : SD_{b_c}) เป็นค่าที่ใช้วัดการกระจายของสัมประสิทธิ์การถดถอยตัวที่ c ถ้าค่าส่วนเบี่ยงเบนมาตรฐานมีค่าน้อย แสดงว่าข้อมูลกระจายรอบค่าเฉลี่ยน้อยหรือมีค่าใกล้เคียงกัน

$$SD_{b_c} = \sqrt{\frac{1}{R-1} \sum_{i=1}^R (\bar{b}_{c(i)} - \bar{b}_c)^2} \quad (21)$$

โดยที่ $\bar{b}_{c(i)}$ คือ ค่าเฉลี่ยของ b_c^* รอบการทำซ้ำที่ i เมื่อ $c = 0, 1$ และ $i = 1, 2, \dots, R$, $\bar{b}_c$ คือ ค่าสัมประสิทธิ์การถดถอยเฉลี่ยรวม

2.3 ค่ารากที่สองของคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error : RMSE) เป็นค่าความคลาดเคลื่อนจากค่าที่ประมาณกับค่าสัมประสิทธิ์การถดถอย ซึ่งถ้าค่า RMSE มีค่าน้อย แสดงว่าค่าที่ได้จากการประมาณนั้นใกล้เคียงกับค่าจริง ดังนั้นถ้าหากค่านี้มีค่าเท่ากับศูนย์ หมายความว่าไม่เกิดความคลาดเคลื่อนในการประมาณ

$$RMSE = \sqrt{\frac{1}{R} \sum_{i=1}^R (\beta_c - \bar{b}_{c(i)})^2} \quad (22)$$

โดยที่ β_c คือ ค่าสัมประสิทธิ์การถดถอยตัวที่ c เมื่อ $c = 0, 1$, $\bar{b}_{c(i)}$ คือ ค่าเฉลี่ยของ b_c^* รอบการทำซ้ำที่ i และ R คือ จำนวนรอบการทำซ้ำ เมื่อ $i = 1, 2, \dots, R$

โดยในการเปรียบเทียบประสิทธิภาพของสมการถดถอยจะพิจารณาจากค่ารากที่สองของคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เป็นอันดับแรก ซึ่งถ้าค่าประมาณพารามิเตอร์วิธีใดมีค่า RMSE มีค่าน้อยที่สุด แสดงว่าค่าที่ได้จากวิธีการประมาณนั้นใกล้เคียงกับค่าจริงมากที่สุดและประสิทธิภาพสูงสุด แต่ถ้าค่าประมาณพารามิเตอร์แต่ละวิธีมีค่า RMSE เท่ากันหรือใกล้เคียงกันมาก จะพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอย (SD_{b_c}) ที่มีค่าน้อยกว่า ซึ่งแสดงว่าข้อมูลกระจายรอบค่าเฉลี่ยน้อยหรือมีค่าใกล้เคียงกัน และพิจารณาค่าสัมประสิทธิ์การถดถอยเฉลี่ย ($\bar{b}_c$) ที่มีค่าใกล้เคียงกับค่าพารามิเตอร์ ซึ่งจะทำให้ได้สมการถดถอยที่มีประสิทธิภาพ ส่งผลให้ค่าทำนาย $\hat{Y}_i$ มีความแม่นยำ

Results

ผลการวิเคราะห์แสดงการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นและประสิทธิภาพของสมการถดถอยเชิงเส้น โดยมีพารามิเตอร์ที่ต้องการประมาณค่า คือ β_0 และ β_1 จากการจำลองข้อมูลภายใต้ค่าความคลาดเคลื่อนที่มีการ

แจกแจงแบบปัวซอง ซึ่งในที่นี้จะแสดงเพียงตารางผลลัพธ์ของ $\lambda = -1, 0, 1$ ที่ค่า $\sigma^2 = 1$ และ 2 เนื่องจากมีผลลัพธ์ไปในทิศทางเดียวกับ $\sigma^2 = 5$ และ 10 ตามลำดับ

จาก Table 1 - Table 2 เมื่อพิจารณาช่วงความเชื่อมั่น 95% ของ β_0 และ β_1 ในทุก ๆ ขนาดตัวอย่าง ซึ่งการพิจารณาประสิทธิภาพของช่วงความเชื่อมั่น จะพิจารณาค่า CP เป็นอันดับแรก พบว่าโดยภาพรวมวิธีที่มีค่า CP มากกว่า 0.939 หรือมีค่ามากที่สุดคือวิธีบูตสแตรป์ที่ แต่ค่า CP สำหรับ β_0 กรณี $\lambda = -1$ และ 1 ของทั้งสองวิธีจะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้นและมีค่าน้อยกว่า 0.939 ในทุกกรณี นอกจากนี้พิจารณาค่า AW และค่า SD_w พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า AW และค่า SD_w มีแนวโน้มลดลงทุกกรณี ถึงแม้ว่าค่า AW และค่า SD_w ของวิธีบูตสแตรป์พีจะมีค่าน้อยกว่าวิธีบูตสแตรป์ทีเล็กน้อย แต่ส่งผลทำให้ค่า CP มีค่าน้อยลงด้วย นั่นหมายความว่าช่วงความเชื่อมั่นที่คำนวณได้จากวิธีบูตสแตรป์พีคลุมค่าพารามิเตอร์ได้น้อยกว่าวิธีบูตสแตรป์ที และเมื่อพิจารณาความแตกต่างของความแปรปรวนของความคลาดเคลื่อน (σ^2) พบว่า ค่า CP กรณี $\sigma^2 = 1$ และ $\sigma^2 = 2$ มีค่าใกล้เคียงกัน แต่ค่า AW และค่า SD_w มีค่ามากขึ้นเมื่อ σ^2 เพิ่มขึ้น เช่นเดียวกับกรณี $\sigma^2 = 5$ และ $\sigma^2 = 10$ นั่นหมายความว่า เมื่อ $\sigma^2 = 1, 2, 5, 10$ ให้ค่า CP ไม่แตกต่างกัน แต่กรณี $\sigma^2 = 10$ จะมีค่า AW และค่า SD_w สูงที่สุด

จาก Table 3 - Table 4 เมื่อพิจารณาประสิทธิภาพของสมการถดถอยจากการประมาณค่า β_0 และ β_1 พบว่าในเกือบทุกกรณี วิธีบูตสแตรป์ทีมีค่า RMSE น้อยกว่าวิธีบูตสแตรป์พี ยกเว้นกรณีที่ $\lambda = 0$ และ $n \leq 30$ และเมื่อขนาดตัวอย่างมากขึ้น ค่า RMSE จะมีแนวโน้มลดลง นอกจากนี้ ค่า b_c ที่ได้จากทั้งสองวิธีให้ค่าใกล้เคียงค่าจริง และเมื่อขนาดตัวอย่างมากขึ้น ค่า SD_{b_c} มีแนวโน้มลดลงในทุกกรณี อีกทั้งเมื่อพิจารณาความแตกต่างของความแปรปรวนของความคลาดเคลื่อน (σ^2) พบว่า กรณี $\sigma^2 = 1$ ทั้งสองวิธีจะมีค่า RMSE น้อยกว่า กรณี $\sigma^2 = 2$ ในทำนองเดียวกันเมื่อ σ^2 เพิ่มขึ้น ค่า b_c สำหรับ β_0 จะห่างจากค่าจริงมากขึ้น และค่า SD_{b_c} จะมีค่าเพิ่มขึ้น ยกเว้น ค่า b_c สำหรับ β_1 จะมีค่าใกล้เคียงค่าจริงถึงแม้ว่าค่า σ^2 จะแตกต่างกัน

Table 1 CP, AW and SD_w value of the confidence intervals for β_0 and β_1 when $X \sim N(3, 2.25)$ and $\epsilon \sim SN(0, 1, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			CP	AW	SD_w	CP	AW	SD_w
-1	20	Bootstrap-p	0.691	1.613	0.399	0.907	0.485	0.117
		Bootstrap-t	<u>0.764</u>	1.808	0.449	<u>0.950</u>	0.696	0.171
	30	Bootstrap-p	0.602	1.324	0.248	0.936	0.397	0.073
		Bootstrap-t	<u>0.708</u>	1.432	0.291	<u>0.974</u>	0.561	0.119
	50	Bootstrap-p	0.400	1.023	0.151	0.922	0.306	0.047
		Bootstrap-t	<u>0.460</u>	1.058	0.172	<u>0.968</u>	0.422	0.070
	100	Bootstrap-p	<u>0.166</u>	0.723	0.077	0.955	0.216	0.024
		Bootstrap-t	0.122	0.723	0.091	<u>0.963</u>	0.290	0.036
0	20	Bootstrap-p	0.914	1.939	0.499	0.915	0.586	0.151
		Bootstrap-t	<u>0.966</u>	2.451	0.588	<u>0.954</u>	0.710	0.178
	30	Bootstrap-p	0.934	1.589	0.290	0.952	0.479	0.088
		Bootstrap-t	<u>0.965</u>	1.865	0.369	<u>0.967</u>	0.549	0.112

Table 1 CP, AW and SD_W value of the confidence intervals for β_0 and β_1 when $X \sim N(3, 2.25)$ and $\varepsilon \sim SN(0, 1, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			CP	AW	SD_W	CP	AW	SD_W
1	50	Bootstrap-p	0.939	1.246	0.181	0.948	0.374	0.055
		Bootstrap-t	<u>0.972</u>	1.417	0.234	<u>0.970</u>	0.418	0.071
	100	Bootstrap-p	0.951	0.876	0.088	0.943	0.261	0.027
		Bootstrap-t	<u>0.962</u>	0.972	0.119	<u>0.957</u>	0.289	0.036
	20	Bootstrap-p	0.667	1.607	0.386	0.901	0.483	0.114
		Bootstrap-t	<u>0.887</u>	2.866	0.702	<u>0.960</u>	0.700	0.180
	30	Bootstrap-p	0.583	1.318	0.249	0.940	0.396	0.075
		Bootstrap-t	<u>0.848</u>	2.288	0.477	<u>0.954</u>	0.562	0.122
	50	Bootstrap-p	0.397	1.026	0.154	0.941	0.307	0.048
		Bootstrap-t	<u>0.744</u>	1.705	0.273	<u>0.962</u>	0.420	0.069
	100	Bootstrap-p	0.138	0.724	0.073	0.939	0.217	0.022
		Bootstrap-t	<u>0.552</u>	1.175	0.141	<u>0.959</u>	0.290	0.035

Note: The confidence level is 0.95.

—— Referring to a CP value that is not less than 0.939 and/or is the highest.

Table 2 CP, AW and SD_W value of the confidence intervals for β_0 and β_1 when $X \sim N(3, 2.25)$ and $\varepsilon \sim SN(0, 2, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			CP	AW	SD_W	CP	AW	SD_W
-1	20	Bootstrap-p	0.678	2.281	0.555	0.915	0.691	0.172
		Bootstrap-t	<u>0.769</u>	2.549	0.663	<u>0.970</u>	0.990	0.258
	30	Bootstrap-p	0.585	1.852	0.355	0.937	0.557	0.106
		Bootstrap-t	<u>0.651</u>	1.997	0.403	<u>0.959</u>	0.791	0.164
	50	Bootstrap-p	0.405	1.443	0.212	0.947	0.434	0.066
		Bootstrap-t	<u>0.451</u>	1.485	0.247	<u>0.957</u>	0.591	0.098
	100	Bootstrap-p	0.128	1.022	0.111	0.939	0.305	0.033
		Bootstrap-t	<u>0.129</u>	1.025	0.127	<u>0.965</u>	0.412	0.052
0	20	Bootstrap-p	0.914	2.750	0.684	0.915	0.826	0.208
		Bootstrap-t	<u>0.957</u>	3.431	0.850	<u>0.954</u>	0.993	0.253
	30	Bootstrap-p	0.947	2.262	0.433	0.947	0.676	0.136
		Bootstrap-t	<u>0.959</u>	2.658	0.540	<u>0.959</u>	0.779	0.158
	50	Bootstrap-p	0.928	1.745	0.268	0.929	0.523	0.081
		Bootstrap-t	<u>0.965</u>	1.992	0.324	<u>0.962</u>	0.591	0.097
	100	Bootstrap-p	0.940	1.242	0.135	0.931	0.371	0.040
		Bootstrap-t	<u>0.967</u>	1.385	0.176	<u>0.970</u>	0.411	0.051
1	20	Bootstrap-p	0.697	2.249	0.546	0.915	0.682	0.173
		Bootstrap-t	<u>0.898</u>	4.061	1.039	<u>0.964</u>	0.992	0.258

Table 2 CP, AW and SD_W value of the confidence intervals for β_0 and β_1 when $X \sim N(3, 2.25)$ and $\varepsilon \sim SN(0, 2, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			CP	AW	SD_W	CP	AW	SD_W
	30	Bootstrap-p	0.607	1.857	0.356	0.936	0.559	0.107
		Bootstrap-t	<u>0.856</u>	3.180	0.656	<u>0.972</u>	0.782	0.165
	50	Bootstrap-p	0.402	1.441	0.217	0.950	0.434	0.066
		Bootstrap-t	<u>0.753</u>	2.409	0.374	<u>0.958</u>	0.595	0.094
	100	Bootstrap-p	0.156	1.023	0.110	0.945	0.306	0.033
		Bootstrap-t	<u>0.557</u>	1.653	0.195	<u>0.965</u>	0.409	0.050

Note: The confidence level is 0.95.

—— Referring to a CP value that is not less than 0.939 and/or is the highest.

Table 3 $\bar{b}_c$, SD_{b_c} and RMSE value of β_0 and β_1 when $X \sim N(3, 2.25)$ and $\varepsilon \sim SN(0, 1, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			$\bar{b}_c$	SD_{b_c}	RMSE	$\bar{b}_c$	SD_{b_c}	RMSE
-1	20	Bootstrap-p	4.449	0.449	0.711	0.997	0.137	0.137
		Bootstrap-t	4.438	0.421	<u>0.496</u>	1.004	0.169	<u>0.119</u>
	30	Bootstrap-p	4.444	0.343	0.653	0.998	0.101	0.101
		Bootstrap-t	4.451	0.314	<u>0.447</u>	0.995	0.125	<u>0.088</u>
	50	Bootstrap-p	4.428	0.282	0.638	1.003	0.084	0.084
		Bootstrap-t	4.447	0.240	<u>0.426</u>	0.998	0.098	<u>0.070</u>
	100	Bootstrap-p	4.447	0.193	0.586	0.997	0.058	0.058
		Bootstrap-t	4.436	0.168	<u>0.416</u>	0.999	0.069	<u>0.049</u>
0	20	Bootstrap-p	5.007	0.530	<u>0.530</u>	1.001	0.162	<u>0.162</u>
		Bootstrap-t	4.973	0.567	0.568	1.006	0.173	0.173
	30	Bootstrap-p	4.999	0.412	<u>0.411</u>	1.002	0.122	<u>0.122</u>
		Bootstrap-t	5.021	0.424	0.425	0.992	0.128	0.128
	50	Bootstrap-p	4.992	0.318	<u>0.318</u>	1.001	0.094	<u>0.094</u>
		Bootstrap-t	5.011	0.322	0.322	0.998	0.097	0.097
	100	Bootstrap-p	5.001	0.226	<u>0.226</u>	1.000	0.067	<u>0.067</u>
		Bootstrap-t	5.000	0.233	0.233	1.000	0.069	0.069
1	20	Bootstrap-p	5.573	0.465	0.738	1.002	0.141	0.141
		Bootstrap-t	5.591	0.650	<u>0.621</u>	0.994	0.158	<u>0.112</u>
	30	Bootstrap-p	5.573	0.345	0.669	0.997	0.102	0.102
		Bootstrap-t	5.553	0.533	<u>0.542</u>	1.006	0.132	<u>0.094</u>
	50	Bootstrap-p	5.572	0.261	0.629	0.999	0.079	0.079
		Bootstrap-t	5.587	0.390	<u>0.498</u>	0.995	0.098	<u>0.069</u>
	100	Bootstrap-p	5.582	0.195	0.614	0.996	0.058	0.058
		Bootstrap-t	5.557	0.281	<u>0.441</u>	1.002	0.070	<u>0.049</u>

Note: —— Referring to a RMSE value that yields the lowest error.

Table 4 $\bar{b}_c$, SD_{b_c} and RMSE value of β_0 and β_1 when $X \sim N(3, 2.25)$ and $\epsilon \sim SN(0, 2, \lambda)$ with $\lambda = -1, 0, 1$

λ	n	Method	β_0			β_1		
			$\bar{b}_c$	SD_{b_c}	RMSE	$\bar{b}_c$	SD_{b_c}	RMSE
-1	20	Bootstrap-p	4.220	0.632	1.004	0.994	0.193	0.193
		Bootstrap-t	4.223	0.564	<u>0.679</u>	0.992	0.229	<u>0.162</u>
	30	Bootstrap-p	4.201	0.506	0.946	1.000	0.149	0.148
		Bootstrap-t	4.192	0.454	<u>0.655</u>	1.005	0.179	<u>0.127</u>
	50	Bootstrap-p	4.205	0.375	0.879	0.999	0.115	0.115
		Bootstrap-t	4.212	0.347	<u>0.609</u>	0.993	0.139	<u>0.098</u>
	100	Bootstrap-p	4.192	0.264	0.850	1.002	0.078	0.078
		Bootstrap-t	4.201	0.234	<u>0.589</u>	1.000	0.095	<u>0.067</u>
0	20	Bootstrap-p	4.973	0.764	<u>0.764</u>	1.007	0.229	<u>0.229</u>
		Bootstrap-t	4.995	0.772	0.771	1.001	0.231	0.231
	30	Bootstrap-p	5.000	0.585	<u>0.585</u>	1.001	0.175	<u>0.174</u>
		Bootstrap-t	4.997	0.610	0.610	1.004	0.184	0.184
	50	Bootstrap-p	5.006	0.470	0.470	0.999	0.139	0.139
		Bootstrap-t	5.000	0.451	<u>0.450</u>	1.000	0.137	<u>0.137</u>
	100	Bootstrap-p	4.984	0.329	0.329	1.005	0.100	0.100
		Bootstrap-t	5.005	0.319	<u>0.319</u>	1.001	0.095	<u>0.095</u>
1	20	Bootstrap-p	5.780	0.592	0.979	1.004	0.177	0.177
		Bootstrap-t	5.803	0.914	<u>0.860</u>	0.998	0.230	<u>0.163</u>
	30	Bootstrap-p	5.767	0.495	0.913	1.007	0.148	0.148
		Bootstrap-t	5.789	0.693	<u>0.742</u>	1.001	0.173	<u>0.123</u>
	50	Bootstrap-p	5.804	0.371	0.885	0.998	0.109	0.109
		Bootstrap-t	5.789	0.563	<u>0.685</u>	1.001	0.141	<u>0.099</u>
	100	Bootstrap-p	5.784	0.262	0.826	1.002	0.078	0.078
		Bootstrap-t	5.774	0.383	<u>0.611</u>	1.006	0.096	<u>0.068</u>

Note: — Referring to a RMSE value that yields the lowest error.

ตัวอย่างการประยุกต์ใช้กับข้อมูลจริง

จากการศึกษาประสิทธิภาพของตัวแบบการถดถอยภายใต้วิธีบูตสแตรป์สองวิธีข้างต้น ต่อไปจะนำมาประยุกต์ใช้กับข้อมูลจริงและเปรียบเทียบประสิทธิภาพของสมการถดถอยโดยจะพิจารณาจากเกณฑ์ค่าสัมประสิทธิ์การกำหนด (Coefficient of Determination : R^2) เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's Information Criterion : AIC) เกณฑ์สารสนเทศของเบส์ (Bayesian Information Criteria : BIC) และค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error : MSE) ที่ต่ำที่สุด เพื่อให้ได้ตัวแบบที่ให้ค่าพยากรณ์ถูกต้องและเหมาะสมที่สุด

ข้อมูลชุดที่ 1 ข้อมูลน้ำหนักทารกแรกเกิด จำนวน 42 ค่าสังเกต โดยกำหนดให้ ตัวแปรอิสระ (X) คือ อายุครรภ์ (สัปดาห์) และตัวแปรตาม (Y) คือ น้ำหนักทารกแรกเกิด (ปอนด์) มีค่าเฉลี่ยของอายุครรภ์ 40 สัปดาห์ ค่าเฉลี่ยของน้ำหนักทารกแรกเกิด 7.263 ปอนด์ ค่าความแปรปรวนของอายุครรภ์เท่ากับ 1.940 สัปดาห์² และค่าความแปรปรวนของน้ำหนัก

ทารกแรกเกิดเท่ากับ 6.985 ปอนด์² การตรวจสอบลักษณะการแจกแจงของข้อมูลน้ำหนักทารกแรกเกิด พบว่า ค่า skewness เท่ากับ -0.056 นั่นคือข้อมูลมีการแจกแจงแบบสมมาตร

Table 5 b_0 , b_1 and 95% confidence interval (U, L) of β_0 and β_1

Method	β_0		β_1	
	b_0	(U, L)	b_1	(U, L)
Bootstrap-p	-6.639	(-10.875,-2.224)	0.355	(0.244,0.465)
Bootstrap-t	-6.720	(-11.002,-1.804)	0.357	(0.231,0.468)

จาก Table 5 ช่วงความเชื่อมั่น 95% สำหรับ β_1 ภายใต้วิธีบูตสแตรป์ทั้งสองวิธี พบว่าช่วงความเชื่อมั่นไม่คลุมค่า ศูนย์และมีความกว้างของช่วงใกล้เคียงกัน แสดงว่าตัวแปรอิสระมีอิทธิพลต่อตัวแปรตาม ที่ระดับนัยสำคัญ 0.05

สมการถดถอยบูตสแตรป์พี คือ $\hat{Y} = -6.639 + 0.355X$

สมการถดถอยบูตสแตรป์ที คือ $\hat{Y} = -6.720 + 0.357X$

Table 6 R^2 , MSE, AIC and BIC of regression model based on bootstrap-p and bootstrap-t methods

Method	R^2	MSE	AIC	BIC
Bootstrap-p	0.482	0.86516	119.563	124.776
Bootstrap-t	<u>0.610</u>	0.86507*	116.802*	122.015*

—— Referring to the highest value of R^2 and * is the shortest value of MSE, AIC and BIC.

จาก Table 6 พบว่าสมการถดถอยที่ให้ค่า R^2 มากที่สุดและให้ค่า MSE AIC และ BIC ต่ำที่สุดคือสมการ ถดถอยบูตสแตรป์ที นั่นคือวิธีที่สร้างสมการถดถอยได้มีประสิทธิภาพพุงสูงที่สุดคือวิธีบูตสแตรป์ที ซึ่งสอดคล้องกับผลลัพธ์ที่ได้จากการจำลองข้อมูลภายใต้ค่าคลาดเคลื่อนมีการแจกแจงเบ้ปกติที่มี $\lambda = 0$ หรือนั่นคือ มีการแจกแจงแบบสมมาตร เมื่อขนาดตัวอย่างเท่ากับ 50 ดังนั้น สมการถดถอยที่ดีที่สุดคือสมการ $\hat{Y} = -6.720 + 0.357X$

ข้อมูลชุดที่ 2 ข้อมูลคะแนนแคลคูลัสและเกรดเฉลี่ยของนักเรียนโรงเรียนแห่งหนึ่ง จำนวน 105 ตัวอย่าง โดย กำหนดให้ตัวแปรอิสระ (X) คือ คะแนนแคลคูลัส และตัวแปรตาม (Y) คือ เกรดเฉลี่ย (GPA) พบว่าค่าเฉลี่ยของคะแนน แคลคูลัสเท่ากับ 78.943 คะแนน ค่าเฉลี่ยของเกรดเฉลี่ยเท่ากับ 82.957 คะแนน ค่าความแปรปรวนของคะแนนแคลคูลัส เท่ากับ 224.910 คะแนน² และค่าความแปรปรวนของเกรดเฉลี่ยเท่ากับ 36.639 คะแนน² การตรวจสอบลักษณะการแจกแจงของข้อมูลเกรดเฉลี่ย พบว่า ค่า skewness เท่ากับ -0.899 นั่นคือ ข้อมูลมีการแจกแจงเบ้ซ้าย (Negatively skewed distribution)

Table 7 b_0 , b_1 and 95% confidence interval (U, L) of β_0 and β_1

Method	β_0		β_1	
	b_0	(U, L)	b_1	(U, L)
Bootstrap-p	-86.178	(-110.925,-63.419)	1.991	(1.722,2.282)
Bootstrap-t	-85.732	(-114.658,-60.053)	1.985	(1.677,2.338)

จาก Table 7 ช่วงความเชื่อมั่น 95% สำหรับ β_1 ภายใต้วิธีบูตสทราปทั้งสองวิธี พบว่า ช่วงความเชื่อมั่นไม่คลุมค่าศูนย์และมีความกว้างของช่วงใกล้เคียงกัน แสดงว่าตัวแปรอิสระมีอิทธิพลต่อตัวแปรตาม ที่ระดับนัยสำคัญ 0.05

สมการถดถอยบูตสทราปพี คือ $\hat{Y} = -86.178 + 1.991X$

สมการถดถอยบูตสทราปที คือ $\hat{Y} = -85.732 + 1.985X$

Table 8 R^2 , MSE, AIC and BIC of regression model based on bootstrap-p and bootstrap-t methods

Method	R^2	MSE	AIC	BIC
Bootstrap-p	0.614	78.819*	773.041	781.002
Bootstrap-t	<u>0.732</u>	78.822	729.635*	737.597*

—— Referring to the highest value of R^2 and * is the shortest value of MSE, AIC and BIC.

จาก Table 8 พบว่าสมการถดถอยที่ให้ค่า R^2 มากที่สุด และให้ค่า AIC และ BIC ต่ำที่สุดคือวิธีบูตสทราปที และทั้งสองวิธีจะให้ค่า MSE ที่ใกล้เคียงกัน ซึ่งสอดคล้องกับผลลัพธ์ที่ได้จากการจำลองข้อมูลภายใต้ค่าคลาดเคลื่อนมีการแจกแจงแบบปกติที่มี $\lambda = -1$ เมื่อขนาดตัวอย่างเท่ากับ 100 นั่นคือโดยภาพรวมสมการถดถอยบูตสทราปทีให้ประสิทธิภาพของสมการถดถอยสูงที่สุด ดังนั้นสมการถดถอยที่ดีที่สุด คือสมการ $\hat{Y} = -85.732 + 1.985X$

Discussion

จากการศึกษาจำลองข้อมูลเชิงสถิติ พบว่าวิธีบูตสทราปที (Bootstrap-t Method) มีประสิทธิภาพในการประมาณช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยสูงกว่าวิธีบูตสทราปพี (Bootstrap-p Method) เกือบทุกขนาดตัวอย่าง แต่ค่า CP ของ β_0 กรณีค่าความคลาดเคลื่อนมีการแจกแจงแบบเบ้ซ้าย ($\lambda = -1$) และเบ้ขวา ($\lambda = 1$) จะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น ซึ่งอาจเป็นผลจากกระบวนการสุ่มซ้ำของวิธีบูตสทราป โดยที่เมื่อขนาดตัวอย่างเพิ่มขึ้นทำให้ข้อมูลที่เกิดจากการสุ่มตัวอย่างซ้ำแบบแทนที่นั้นจะเข้าใกล้การแจกแจงสมมาตรแทนที่จะมีการแจกแจงเบ้ตามที่ควรจะเป็น แต่อย่างไรก็ตามเมื่อความแปรปรวนของความคลาดเคลื่อน (σ^2) มีค่าแตกต่างกันหรือเพิ่มสูงขึ้นไม่ส่งผลต่อประสิทธิภาพของค่า CP แต่ค่า AW และค่า SD_w สูงขึ้นเมื่อ σ^2 เพิ่มขึ้น โดยภาพรวมการประมาณช่วงความเชื่อมั่นของ β_0 และ β_1 ที่ได้จากวิธีบูตสทราปทีมีประสิทธิภาพสูงกว่าวิธีบูตสทราปพี ซึ่งสอดคล้องกับงานวิจัยของ Siti Fairus Mokhtar และคณะ (Mokhtar et al., 2023) ที่ว่าวิธีบูตสทราปทีมีความแม่นยำและประสิทธิภาพดีกว่าวิธีบูตสทราปพี โดยเฉพาะในกรณีที่ข้อมูลมีการเบี่ยงเบนจากการแจกแจงปกติ หรือเมื่อมีความไม่แน่นอนสูงในการประมาณค่า เนื่องจากบูตสทราปทีสามารถปรับค่าความแปรปรวนได้ในแต่ละรอบของการสุ่มซ้ำ ทำให้ช่วงความเชื่อมั่นที่ได้มีความน่าเชื่อถือมากกว่า

นอกจากนี้พิจารณาประสิทธิภาพของสมการถดถอยจากการประมาณพารามิเตอร์ พบว่าเกือบทุกขนาดตัวอย่างวิธีบูตสทราปทีมีค่า RMSE น้อยกว่าวิธีบูตสทราปพี ยกเว้นกรณีที่ค่าความคลาดเคลื่อนมีการแจกแจงแบบสมมาตร ($\lambda = 0$) ที่ขนาดตัวอย่างเล็ก ($n \leq 30$) และเมื่อขนาดตัวอย่างมากขึ้น ค่า RMSE จะมีแนวโน้มลดลง นอกจากนี้ถ้าความแปรปรวนของความคลาดเคลื่อน (σ^2) มีค่าเพิ่มขึ้น จะได้ค่า RMSE และ ค่า SD_{b_c} เพิ่มขึ้น โดยที่ค่า b_c สำหรับ β_0 จะห่างจากค่าจริงมากขึ้น แต่ค่า b_c สำหรับ β_1 จะมีค่าใกล้เคียงค่าจริงเช่นเดิม โดยภาพรวมการประมาณค่าพารามิเตอร์ด้วยวิธีบูตสทราปทีจะให้ค่าที่แม่นยำกว่าวิธีบูตสทราปพี หมายความว่าสมการถดถอยภายใต้วิธี

สูตรที่ 1 ที่มีประสิทธิภาพมากกว่าสมการถดถอยภายใต้วิธีสูตรที่ 1 ซึ่งสอดคล้องกับงานวิจัยของ Mahendra Saha และคณะ (Saha *et al.*, 2018) วิธีสูตรที่ 1 ที่มีความแม่นยำทางทฤษฎีดีกว่า โดยเฉพาะเมื่อขนาดตัวอย่างใหญ่ และตัวประมาณมีการเบี่ยงเบน แต่ในกรณีที่ขนาดตัวอย่างเล็กหรือข้อมูลมีลักษณะเฉพาะ วิธีสูตรที่ 1 อาจให้ผลลัพธ์ที่เสถียรกว่า

Conclusions

การเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นและสมการถดถอยเชิงเส้นภายใต้วิธีสูตรที่ 1 แบบไม่อิงพารามิเตอร์ 2 วิธี ได้แก่ วิธีสูตรที่ 1 และวิธีสูตรที่ 2 จากการจำลองข้อมูลภายใต้ค่าความคลาดเคลื่อนที่มีการแจกแจงแบบเบ้ปกติ 3 ลักษณะ ได้แก่ เบ้ซ้าย เบ้ขวา และสมมาตร พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ช่วงความเชื่อมั่นที่คำนวณจากวิธีสูตรที่ 1 ทั้งสองวิธีมีความน่าจะเป็นที่จะครอบคลุมค่าพารามิเตอร์เพิ่มขึ้นถ้าข้อมูลแจกแจงแบบสมมาตร แต่จะมีความน่าจะเป็นที่จะครอบคลุมค่าพารามิเตอร์ลดลงถ้าข้อมูลมีการแจกแจงแบบเบ้ แต่อย่างไรก็ตามวิธีสูตรที่ 1 จะมีประสิทธิภาพในการประมาณช่วงความเชื่อมั่นสูงกว่าวิธีสูตรที่ 2 เนื่องจากมีความน่าจะเป็นที่จะครอบคลุมค่าพารามิเตอร์ได้สูงกว่าเกือบทุกกรณี และถ้าความแปรปรวนของความคลาดเคลื่อนเพิ่มขึ้น ค่าความน่าจะเป็นที่จะครอบคลุมค่าพารามิเตอร์จะไม่เพิ่มขึ้น แต่จะมีความกว้างเฉลี่ยของช่วงความเชื่อมั่นและส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงเพิ่มขึ้น

ผลจากการเปรียบเทียบประสิทธิภาพของสมการถดถอยภายใต้วิธีสูตรที่ 1 พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น สูตรที่ 1 ทั้งสองวิธีสามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงมากยิ่งขึ้น เนื่องจากมีความคลาดเคลื่อนระหว่างค่าประมาณกับค่าสัมประสิทธิ์การถดถอยจริงน้อยลง และค่าประมาณที่ได้จากวิธีสูตรที่ 1 ที่มีความคลาดเคลื่อนจากการประมาณน้อยกว่าวิธีสูตรที่ 2 เกือบทุกกรณี ยกเว้นกรณีที่ค่าความคลาดเคลื่อนมีการแจกแจงแบบสมมาตรและขนาดตัวอย่างเล็ก นอกจากนี้ถ้าความแปรปรวนของความคลาดเคลื่อนมีค่าเพิ่มขึ้น ค่าประมาณที่ได้จากวิธีสูตรที่ 1 ทั้งสองวิธีจะมีความคลาดเคลื่อนและส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์การถดถอยเพิ่มขึ้น โดยภาพรวมการประมาณค่าพารามิเตอร์ด้วยวิธีสูตรที่ 1 จะมีประสิทธิภาพกว่าวิธีสูตรที่ 2 หมายความว่าสมการถดถอยภายใต้วิธีสูตรที่ 1 ที่มีประสิทธิภาพมากกว่าสมการถดถอยภายใต้วิธีสูตรที่ 2

ผลลัพธ์ที่ได้จากการประยุกต์ใช้กับข้อมูลจริงนั้นสอดคล้องกับการจำลองข้อมูลเชิงสถิติภายใต้การแจกแจงแบบเบ้ปกติ ดังนั้น วิธีสูตรที่ 1 จึงเป็นวิธีทางเลือกในการประมาณค่าช่วงความเชื่อมั่นและการประมาณค่าสัมประสิทธิ์การถดถอย โดยเฉพาะวิธีสูตรที่ 1 ซึ่งส่งผลให้สมการถดถอยที่ได้นั้นมีประสิทธิภาพในการพยากรณ์ได้อย่างแม่นยำ

Acknowledgments

งานวิจัยนี้สำเร็จสมบูรณ์ได้ด้วยดีเพราะได้รับความอนุเคราะห์จากบุคคลหลาย ๆ ท่าน รวมถึงมหาวิทยาลัยศรีนครินทรวิโรฒที่ได้อำนวยความสะดวกในเรื่องของสถานที่และอุปกรณ์ในการทำงานวิจัย คณะผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

References

- Davison, A.C., & Kuonen, D. (2002). An Introduction to the Bootstrap with Applications in R. *Statistical Computing & Statistical Graphics Newsletter*, 13(1), 6-11.
- Dikta, G., & Scheer, M. (2021). *Bootstrap methods: with applications in R*. Springer Nature.
- Eck, D. J. (2018). Bootstrapping for multivariate linear regression models. *Statistics & Probability Letters*, 134, 141-149.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26.
- Efron, B. (1982). *The Jackknife, The Bootstrap and Other Resampling Plans*. Society for industrial and applied mathematics.
- Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. (1st ed.). New York: Chapman & Hall/CRC.
- Figueiredo, F., & Gomes, M.I. (2012). The Skew-Normal Distribution in SPC. *REVSTAT-Statistical Journal*, 11(1), 83-104.
- Hall, P. (1988). Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics*, 16(3), 927-953.
- Kashif, M., Aslam, M., Rao, G. S., AL-Marshadi, A. H., & Jun, C. H. (2017). Bootstrap Confidence Intervals of the Modified Process Capability Index for Weibull Distribution. *Arabian Journal for Science and Engineering*, 42, 4565-4573.
- Kokaew, A., Thaithanan, J., Bodhisuwan, W., & Volodin, A. (2021). Confidence Estimation of a Ratio of Binomial Proportions for Dependent Populations. *Lobachevskii Journal of Mathematics*, 42(2), 394-403.
- Mokhtar, S. F., Yusof, Z. M., & Sapiri, H. (2023). Confidence intervals by bootstrapping approach: a significance review. *Malaysian Journal of Fundamental and Applied Sciences*, 19(1), 30-42.
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, Retrieved from <https://www.r-project.org/>.
- Saha, M., Dey, S., & Maiti, S. S. (2018). Parametric and non-parametric bootstrap confidence intervals of C_{Npk} for exponential power distribution. *Journal of Industrial and Production Engineering*, 35(3), 160-169.

Thiuthad, P. (2021). The Univariate Skew-Normal Distribution. *The Journal of Applied Science*, 20(2), 293-304.

(in Thai)

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing*. (4th ed.). San Diego, CA:
Academic Press.