บทคัดย่อ

ในการศึกษาและการวิจัยในทางสถิติต่างๆ มีโอกาสที่จะละทิ้งตัวแปรอิสระสำคัญออกไปจากแบบจำลอง ซึ่งทำให้เกิดปัญหาความคลาดเคลื่อนจากการระบุแบบจำลองไม่ถูกต้อง (specification error) และส่งผลให้ค่าพารามิเตอร์ที่ประมาณค่าได้จากแบบจำลองที่ผิดพลาดนั้นมีความเอนเอียง (biased) ในกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์ (r) ระหว่างตัวแปรอิสระที่อยู่ในแบบจำลองกับตัวแปรที่ละทิ้งออกไปจากแบบจำลองมีค่ามาก จะทำให้เกิดการเปลี่ยนแปลงอย่างมากในค่าของพารามิเตอร์ที่ประมาณมาได้ การศึกษาครั้งนี้มุ่งเน้นวิเคราะห์ผลของความเอนเอียง (biased) ที่เกิดขึ้นกับการประมาณค่าพารามิเตอร์ค่า R^2 และค่าความแปรปรวน ($\hat{\sigma}^2$) ในแบบจำลองที่มีการละทิ้งตัวแปรอิสระ

การศึกษานี้ ใช้วิธีการทดลองโดยสร้างแบบจำลองที่แท้จริงด้วยการสร้างข้อมูลจากวิธี Monte Carlo และสร้างแบบจำลองทดสอบ ซึ่งมีตัวแปรอิสระ 4 ตัว (X_1-X_4) โดยให้ตัวแปร X_4 เป็นตัวแปรที่ถูกละทิ้ง ค่าสัมประสิทธิ์สหสัมพันธ์ (r) ระหว่าง X_4 และ X_2 กับ X_3 มีค่าอยู่ระหว่าง 0-0.3 และให้ความสัมพันธ์ระหว่าง X_1 กับ X_4 มีช่วง คือ 0-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9, 0.9-1.0

ผลการศึกษาการประมาณค่าความเอนเอียงของค่าสัมประสิทธิ์ พบว่า ค่าเฉลี่ยของการเอนเอียงในการประมาณค่าสัมประสิทธิ์ในตัวแปร X_1 ที่กำหนดให้มีความสัมพันธ์กับตัวแปรที่ละทิ้งออกไป (X_4) จะมีค่าสูงมากเมื่อค่า r มีค่าเข้าใกล้ 1 และสำหรับค่าคงที่นั้นก็จะมีค่าเฉลี่ยของความเอนเอียงในการประมาณค่าสัมประสิทธิ์ที่สูงมากไม่ว่าค่า r จะเป็นเท่าไรก็ตาม ส่วนตัวแปรอื่นๆ (X_2 กับ X_3) ที่มีความสัมพันธ์ไม่มากกับตัวแปรที่ละทิ้งออกไป (X_4) ก็มีค่าเฉลี่ยของความเอนเอียงในการประมาณค่าสัมประสิทธิ์ในระดับค่อนข้างสูงแต่ไม่เท่ากับค่าสัมประสิทธิ์ของตัวแปร X_1

ในการหาอัตราส่วนของความเอนเอียงในการประมาณค่าสัมประสิทธิ์อันเนื่องมาจากค่า r พบว่า เมื่อค่า r มีค่า 0-0.7 ปัจจัยที่สำคัญที่สุดที่มีผลต่อความเอนเอียงในการประมาณค่าสัมประสิทธิ์ทุกตัวแปรยกเว้นตัวแปร X_1 คือ ปัจจัยที่มาจากการละทิ้งตัวแปรที่สำคัญออกไป รองลงมาก็คือปัจจัยที่มาจากความสัมพันธ์ระหว่างตัวแปรที่กำหนดให้มีความสัมพันธ์กับตัวแปรที่ละทิ้งออกไป ($r_{x_1x_4}$) สำหรับผลกระทบที่มีต่อตัวแปร X_1 นั้นได้แก่ปัจจัย $r_{x_1x_4}$ เพียงอย่างเดียวที่เป็นปัจจัยสำคัญ แต่เมื่อค่า r มีค่ามากกว่า 0.7 ขึ้นไปแล้ว ปัจจัยที่สำคัญที่สุดต่อความเอนเอียงในการประมาณค่าสัมประสิทธิ์ทุกตัวแปรคือ ปัจจัยที่มาจากการละทิ้งตัวแปรที่สำคัญออกไป รองลงมาก็คือปัจจัยที่มาจากค่า $r_{x_1x_4}$ ซึ่งค่าอัตราส่วนที่ได้มานี้ก็ยังสอดคล้องกับผลของการถดถอยที่แสดงว่าค่าคงที่หรือปัจจัยที่มาจากการละทิ้งตัวแปรที่สำคัญออกไปกับปัจจัยที่มาจากค่า $r_{x_1x_4}$ เป็นปัจจัยที่มีนัยสำคัญทางสถิติ ส่วนปัจจัย $r_{x_2x_4}$, $r_{x_3x_4}$ จะไม่มีนัยสำคัญทางสถิติ

สำหรับการประมาณค่าความเอนเอียงของค่าสัมประสิทธิ์ของการตัดสินใจ (R^2) พบว่า การลดลงของค่า R^2 จะเกิดจากการละทิ้งตัวแปรอิสระที่เกี่ยวข้องมากที่สุด ไม่ว่าค่า r จะมีค่าเท่าไรก็ตาม แต่ค่า r ที่สูงขึ้นนี้จะทำให้การเปลี่ยนแปลงในค่า R^2 มีค่าลดลง

สำหรับการประมาณค่าความแปรปรวนของความคลาดเคลื่อน ($\hat{\sigma}^2$) พบว่า มีค่าเพิ่มขึ้นเป็นอย่างมากและยังก่อให้เกิดปัญหา heteroscedasticity ด้วย และยังทำให้ความแปรปรวนของค่าสัมประสิทธิ์ของทุกตัวแปร ($\text{var}(\hat{\beta})$) มีค่าเพิ่มขึ้น โดยเฉพาะในค่าคงที่ แต่การเพิ่มขึ้นของค่า r ไม่ได้มีผลต่อการเปลี่ยนแปลงแต่อย่างใด ส่วนค่าสถิติ Durbin-Watson จะไม่มีการเปลี่ยนแปลงแต่อย่างใด

ABSTRACT

In statistical studies, some relevant variables might be omitted which cause model specification error and consequently bias in estimated parameters. The higher the value of simple correlation (r) between explanatory variables in the model and the omitted relevant variables, the greater the bias in the estimated parameters is expected.This research, therefore, aimed to detect the bias in estimated parameters and bias in R^2 and $\hat{\sigma}^2$ when a relevant variable was omitted.

The Monte Carlo method was employed to create the pseudorandom numbers under various ranges of r i.e. 0-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9, 0.9-1.

The findings indicated that the average bias of the estimated coefficient X_1 due to the omitted variable (X_4) would be very high when the simple correlation approached 1. For the constant or intercept term, the average bias of the estimated coefficient was extremely high no matter what the r value would be. For other variables (X_2 and X_3) which were not much related to the omitted variable (X_4), their average bias of the estimated coefficient were also very high but not as much as those of X_1 and the constant term.

In addition, the study of the ratio of the biased estimation in estimated coefficient and the simple correlation showed that when simple correlation had value between 0-0.7, omitting relevant variable was the most important factor influencing the bias in estimated coefficient in every variable except the determined variable (X_1). The second important factor was the indirect influence of $r_{x_1x_4}$. For X_1, the only influential factor was $r_{x_1x_4}$:

However, when simple correlation had value greater than 0.7, the most important factor causing the bias in estimated coefficient in every variable was the omitted variable. The second important factor was $r_{x_1x_4}$. And biases were investigated by regressing biases and simple correlations. The results indicated that the constant term (or the factor resulted from the variable omission) and $r_{x_1x_4}$ were statistically significant factors but $r_{x_2x_4}$, $r_{x_3x_4}$ were insignificant factors.

For the biased estimation of the coefficient of determinant (R^2), the result indicated that the decreasing of R^2 value was most influenced by the omission of X_4 regardless of r values. However, the change in R^2 value decreased with the increase in r value.

For the estimation of the variance (σ^2), the result showed that it was highly increased and also constituted the heteroscedasticity. Consequently, this caused the variance of all coefficients, especially in constant term, to increase. However, the change in r value did not affect these outcomes. There was no change for Durbin-Watson statistic.