กัญญารัตน์ โพธิสุทธิ์ : การประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้นเมื่อมีค่าผิดปกติในตัวแปรตาม (ESTIMATION OF PARAMETERS IN LINEAR REGRESSION MODEL HAVING OUTLIERS IN DEPENDENT VARIABLE). อ. ที่ปรึกษา : รศ.ร.อ.มานพ วราภักดิ์, 150 หน้า. ISBN 974-53-1934-1

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณในการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้น เมื่อมีค่าผิดปกติในตัวแปรตาม โดยทำการเปรียบเทียบตัวประมาณกำลังสองน้อยที่สุดแบบสามัญ (OLSE- Ordinary Least Squares Estimator) ตัวประมาณกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักที่ได้รับการปรับ (AWLSE - Adaptive Weighted Least Squares Estimator) และตัวประมาณกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักที่มีความแกร่งและมีประสิทธิภาพ (REWLSE - Robust and Efficient Weighted Least Squares Estimator) ซึ่งเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของตัวประมาณคือ ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) ของพารามิเตอร์ สถานการณ์ที่ศึกษาคือกำหนดการแจกแจงความคลาดเคลื่อนสุ่ม($\underline{\varepsilon}$) สองการแจกแจง คือการแจกแจงแบบปกติปลอมปนระหว่าง N(0,10) กับ $N(0,10C^2)$โดยกำหนดให้สเกลแฟกเตอร์ (C) มีค่าเท่ากับ 3 สำหรับข้อมูลที่มีค่าผิดปกติในระดับไม่รุนแรง และสเกลแฟกเตอร์เท่ากับ 12 สำหรับข้อมูลที่มีค่าผิดปกติในระดับรุนแรง และจากการแจกแจงแบบปกติปลอมปนระหว่าง N(0,10) กับ $L(0,\beta)$ โดยกำหนดให้ β =8 เมื่อข้อมูลมีค่าผิดปกติในระดับไม่รุนแรง และ $\beta = 25$ สำหรับข้อมูลที่มีค่าผิดปกติในระดับรุนแรง กำหนดค่าพารามิเตอร์ $\underline{\beta} = (5, 1, 1)^T$ ตัวแปรอิสระ x_1 จำลองมาจากการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 20 และความแปรปรวนเท่ากับ 10 ตัวแปรอิสระ x_2 จำลองมาจากการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 30 และความแปรปรวนเท่ากับ 25 โดยแต่ละระดับความรุนแรงของค่าผิดปกติจะกำหนดให้มีขนาดตัวอย่าง (n) เท่ากับ 20, 30, 40, 50, 60, 70, 80, 90 และ 100 และสัดส่วนการปลอมปน (p) เท่ากับ 0.05, 0.10, 0.15 และ 0.20 จำลองสถานการณ์การทดลองด้วยเทคนิคมอนติคาร์โลซึ่งทำซ้ำ 500 ครั้งในแต่ละสถานการณ์

ผลการวิจัยปรากฏว่าระดับค่าผิดปกติ สัดส่วนการปลอมปน และขนาดตัวอย่าง ต่างมีผลต่อตัวประมาณค่าพารามิเตอร์ของทั้ง 3 ตัว โดยค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของพารามิเตอร์จะเพิ่มขึ้นเมื่อระดับค่าผิดปกติหรือสัดส่วนการปลอมปนเพิ่มขึ้น แต่จะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

กรณีที่ไม่มีค่าผิดปกติในตัวแปรตามและในตัวแปรอิสระ

ในทุกขนาดตัวอย่างและทุกสัดส่วนการปลอมปน ตัวประมาณ OLS ให้ประสิทธิภาพในการประมาณสูงที่สุด และเมื่อขนาดตัวอย่างตั้งแต่ 60 ขึ้นไป ตัวประมาณ OLS ตัวประมาณ AWLS และตัวประมาณ REWLS จะมีค่า AMSE ใกล้เคียงกัน

กรณีที่ตัวแปรตามมีค่าผิดปกติในระดับไม่รุนแรง

กรณีที่สัดส่วนการปลอมปนน้อย ($p \in [0.05,0.10]$) และขนาดตัวอย่างมีขนาดเล็ก ($n \in [20,30]$) ตัวประมาณ REWLS ให้ประสิทธิภาพในการประมาณสูงที่สุด ในขณะที่สัดส่วนการปลอมปนน้อย ($p \in [0.05,0.10]$) และขนาดตัวอย่างเพิ่มขึ้น ($n \in (30,100]$) ตัวประมาณ AWLS ให้ประสิทธิภาพในการประมาณสูงที่สุด สำหรับกรณีที่สัดส่วนการปลอมปนเพิ่มขึ้น ($p \in (0.10,0.20]$) ในทุกๆขนาดตัวอย่าง ($n \in [20,100]$) ตัวประมาณ AWLS ให้ประสิทธิภาพในการประมาณสูงที่สุด

กรณีที่ตัวแปรตามมีค่าผิดปกติในระดับรุนแรง

ในทุกขนาดตัวอย่างและทุกสัดส่วนการปลอมปน ตัวประมาณ REWLS ให้ประสิทธิภาพในการประมาณสูงที่สุด และเมื่อสัดส่วนการปลอมปนน้อย (p = 0.05) และขนาดตัวอย่างตั้งแต่ 40 ขึ้นไป พบว่าตัวประมาณ AWLS และตัวประมาณ REWLS มีประสิทธิภาพในการประมาณใกล้เคียงกัน

ภาควิชา..... สถิติ.....
สาขาวิชา..... สถิติ .....
ปีการศึกษา..... 2547 .....

ลายมือชื่อนิสิต กัญญารัตน์ โพธิสุทธิ์
ลายมือชื่ออาจารย์ที่ปรึกษา

4582161426 : MAJOR STATISTICS

KEY WORD : LINEAR REGRESSION / OUTLIERS / ADAPTIVE WEIGHTED LEAST SQUARES ESTIMATOR / ROBUST AND EFFICIENT WEIGHTED LEAST SQUARES ESTIMATOR

KANYARAT POTISUT : ESTIMATION OF PARAMETERS IN LINEAR REGRESSION MODEL HAVING OUTLIERS IN DEPENDENT VARIABLE. THESIS ADVISOR : ASSOC. PROF. CAPT. MANOP VARAPHAKDI , M.S. 150 pp. ISBN 974-53-1934-1

The objective of this research is to compare the efficiency of estimators for parametor estimation in linear regression model when the dependent variable has outliers. The estimators are Ordinary Least Squares Estimator (OLSE), Adaptive Weighted Least squares Estimator (AWLSE), and Robust and Efficient Weighted Least Squares Estimator (REWLSE). The measurement for the efficiency of estimators is the Average Mean Square Error (AMSE). Random Errors ($\underline{\varepsilon}$) are independent and identically distributed normal that are generated from two distributions and was done under mild and extreme outliers. The contaminated normal distribution is mixture of the normal distribution having mean of zero and variance of 10, and the normal distribution having mean of zero and variance of $10C^2$ where C is a scale factor that is 3 for mild level and 10 for extreme level. And the contaminated normal distribution is mixture of the normal distribution having mean of zero and variance of 10 and the Laplace distribution having mean of zero and variance of $2\beta^2$ where β is 8 for mild level and 25 for extreme level. This research specified the parameter $\underline{\beta} = (5, 1, 1)^T$. The observations of independent variable X_1 are generated from the normal distribution with mean of 20 and variance of 10. The observations of independent variable X_2 are generated from the normal distribution with mean of 30 and variance of 25. The sample sizes (n) are 20, 30, 40, 50, 60, 70, 80, 90 and 100. The proportions of contamination (p) are 0.05, 0.10, 0.15 and 0.20. The AMSE of the estimators are computed through the Monte Carlo Simulation method. This simulation is repeated 500 times in each situation.

The results of this research show that the level of outliers, proportions of contamination and sample sizes have effected on the parameter estimations. The average values of mean square error of parameters increase when level of outliers or proportions of contamination increase but they decrease when the sample sizes increase.

In case of no outliers in dependent variable and independent variables

For all sample sizes and proportions of contamination, OLSE is the most efficient. Whereas $n \geq 60$, the AMSE of OLSE, AWLSE and REWLSE are nearly the same.

In case of dependent variable has mild outliers

For small proportions of contamination ($p \in [0.05,0.10]$) and sample sizes ($n \in [20,30]$), REWLSE is the most efficient. Whereas AWLSE is the most efficient when sample size increases ($n \in (30,100]$). For large proportions of contamination ($p \in 0.10,0.20]$) and for all n ($n \in [20,100]$), AWLSE is the most efficient.

In case of dependent variable has extreme outliers

For all sample sizes and proportions of contamination, REWLSE is the most efficient. But the AMSE of AWLSE and REWLSE are a nearly efficiency at $p = 0.05$ for $n \geq 40$

Department.....Statistics.....

Field of study.....Statistics.....

Academic year.....2004.....

Student's signature...กัญญารัตน์ โพธิ์สุทธิ์

Advisor's signature...