บทคัดย่อ T158765

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษา และเปรียบเทียบประสิทธิภาพในการตรวจสอบค่าผิดปกติของตัวสถิติทดสอบ 3 ตัว ได้แก่ (1) ตัวสถิติทดสอบของ Marasinghe (MV) (2) ตัวสถิติทดสอบของ Kainifard and Swallow โดยใช้ตัวประมาณสเกลแบบ biweight เป็นตัวประมาณของค่าเบี่ยงเบนมาตรฐาน (WSB) และ (3) ตัวสถิติทดสอบของ Kainifard and Swallow โดยใช้ค่าพิสัยระหว่างควอไทล์ คำนวณค่าประมาณของค่าเบี่ยงเบนมาตรฐาน (WIR) โดยศึกษาในกรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบหางยาว ได้แก่ การแจกแจงปกติปลอมปนในตำแหน่ง และปลอมปนในสเกล กรณีที่ความคลาดเคลื่อนมีการแจกแจงเบ้ ได้แก่ การแจกแจงลอกนอร์มอล และแกมมา ซึ่งแต่ละการแจกแจงศึกษาในกรณีที่จำนวนตัวแปรอิสระเท่ากับ 1, 3 และ 5 ขนาดตัวอย่างเท่ากับ 20, 50 และ 100 โดยจำนวนค่าผิดปกติ 0, 1, 2 และ 3 ค่า การวิจัยสรุปผลได้ดังนี้

เมื่อความคลาดเคลื่อนมีการแจกแจงแบบหางยาว และการแจกแจงเบ้ พบว่าตัวสถิติทดสอบ MV มีสัดส่วนในการตรวจพบค่าผิดปกติทั้งที่ในชุดข้อมูลนั้นไม่มีค่าผิดปกติอยู่ในเกณฑ์ของ Bradley มากที่สุด รองลงมาคือ ตัวสถิติทดสอบ WSB และตัวสถิติทดสอบ WIR ตามลำดับ ผลการเปรียบเทียบประสิทธิภาพของตัวสถิติทดสอบด้วยสัดส่วนในการตรวจพบค่าผิดปกติจริงทุกค่า สัดส่วนในการเกิดการบดบังค่าผิดปกติ และสัดส่วนในการเกิดการแฝงตัวของค่าผิดปกติ พบว่ากรณีที่มีจำนวนค่าผิดปกติ 1 ค่า ที่ขนาดตัวอย่างเท่ากับ 20 ตัวสถิติทดสอบ MV มีประสิทธิภาพดีที่สุดทุกการแจกแจง แต่เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 เมื่อความคลาดเคลื่อนมีการแจกแจงแบบหางยาว และแกมมา พบว่าตัวสถิติทดสอบ WSB และ WIR มีประสิทธิภาพใกล้เคียงกัน และมีประสิทธิภาพดีกว่าตัวสถิติทดสอบ MV เมื่อความคลาดคลื่อนมีการแจกแจงลอกนอร์มอล พบว่าตัวสถิติทดสอบ WIR มีประสิทธิภาพดีที่สุด รองลงมาคือ ตัวสถิติทดสอบ MV และตัวสถิติทดสอบ WSB ตามลำดับ กรณีที่จำนวนค่าผิดปกติ 2 และ 3 ค่า พบว่า ตัวสถิติทดสอบ MV มีประสิทธิภาพดีที่สุดในทุกการแจกแจง

The objective of this study was to investigate three methods of detecting outliers in multiple linear regression. They are Marasinghe method (MV) and two robust scale estimators by Kainifard and Swallow : biweight scale estimator (WSB) and interquartile range (WIR). Long-tailed distribution were generated by scale and location contamination of normal distributions. The skewed distribution cases were lognormal and gamma distributions. For each distribution considered in this study 1, 3 and 5 independent variables were used and sample sizes of 20, 50 and 100 were used. The number of outliers present were 0, 1, 2 or 3.

In the cases of long-tailed and skewed distributions the MV method had the largest proportion of samples without outliers declared to contain outliers using Bradley's criterion, WSB method was second and WIR had the smallest proportion. Considering the efficiency of the three methods from the proportion of samples in which outliers were correctly detected, proportion with a masking effect and the proportion with a swamping effect. When there was one with outlier it was found that, for all distributions, the MV method had the highest efficiency for a sample size of 20. For sample sizes of 50 and 100 it was found that the WSB and WIR methods had higher efficiencies than the MV method for long-tailed and gamma distributions. But the WIR method had the highest efficiency, the MV method was second and the WSB method was the lowest for lognormal distributions. For two and three outliers the MV method always had the highest efficiency.