The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.399
29(3), 399–408

Robust tests for heteroscedasticity using outlier detection methods

Han Son Seo[a] · Min Yoon[b,1]

[a]Department of Applied Statistics, Konkuk University;
[b]Department of Statistics, Pukyong National University

(Received November 2, 2015; Revised February 10, 2016; Accepted March 12, 2016)

Abstract

There is a need to detect heteroscedasticity in a regression analysis; however, it invalidates the standard inference procedure. The diagnostics on heteroscedasticity may be distorted when both outliers and heteroscedasticity exist. Available heteroscedasticity detection methods in the presence of outliers usually use robust estimators or separating outliers from the data. Several approaches have been suggested to identify outliers in the heteroscedasticity problem. In this article conventional tests on heteroscedasticity are modified by using a sequential outlier detection methods to separate outliers from contaminated data. The performance of the proposed method is compared with original tests by a Monte Carlo study and examples.

Keywords: heteroscedasticity, linear regression model, outliers, robust tests

1. 서론

통계자료분석의 대표적 기법중 하나인 회귀분석은 모형에 관련된 여러 가정을 전제하고 수행된다. 그 중 오차항의 등분산(homoscedasticity) 가정이 성립되지 않는 이분산(heteroscedasticity) 문제가 발생하면 최소자승추정법에 의한 회귀모형 계수 추정치들은 최소분산 속성을 잃게되어 결과적으로 모형의 추정에서 최적의 값을 도출하지 못하거나 검정에서 부정확한 결과를 초래하게 된다. 등분산 가정이 보장되지 않는 사례는 실제 자료에서 쉽게 찾을 수 있다. 예를 들면 규모의 크기에 따라 기업이나 개인의 소득분포는 다를 수 있고 값의 퍼짐정도도 일정하지 않으며, 투약환경이 다른 환자에 대한 임상실험의 결과에서도 등분산이 보장되지 않는 것이 일반적이다 (Carroll과 Ruppert, 1988). 이분산 여부를 판단하기 위해 제안된 방법들은 검정 또는 그림을 활용한 접근법으로 구분된다. 이분산 탐지를 위한 그림의 기본적 형태는 잔차와 추정치 또는 잔차와 설명변수의 산점도이며 이를 응용한 다양한 형태의 그림들이 제시되었다 (Cook과 Weisberg, 1983; Draper와 Smith, 1998; Ryan, 2008, Weisberg, 2005). 그림을 이용한 방법은 관찰치의 정보를 모두 이용하는 반면 주관적 판단에 의존하게 되므로 계량적 검정절차와 상호 보완적으로 사용된다. 이분산 검정은 Anscombe (1961), Bickel (1978), Breusch와 Pagan (1979), Cook과 Weisberg (1983), Goldfeld와 Quandt (1965), Horn (1981), White (1980) 등에 의해 다양하

This paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2015.
[1]Corresponding author: Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: myoon@pknu.ac.kr

게 제안되었다. 이분산 문제는 자료의 본질적 경향이 아니라 모형의 오류나 이상치의 존재에 의해서 왜곡될 수 있으며 특히 이상치를 포함한 자료에서는 이상치와 이분산에 각각 작용하는 관찰치를 구분하는 것이 쉽지 않다. 이상치를 고려한 이분산 검정법은 다양하게 제안되었다 (Bickel, 1978; Carroll과 Ruppert, 1981; Hammerstrom, 1981). Hubert와 Rousseeuw (1997), Giloni 등 (2006)은 가중 최소절대오차(weighted least absolute deviation; WLAD) 추정량에 의한 검정을 제안하였고 Cheng (2012)은 Atkinson (1985)이 제안한 jigsaw 그림을 활용하는 방법을 제안하였으며 Rana 등 (2008)도 기존 검정법에 강건통계량을 적용한 방법을 제안하였다. 본 연구에서는 이상치를 고려한 이분산 검정법으로써 기존의 이분산 검정법 중 Breusch와 Pagan (1979), Goldfeld와 Quandt (1965), White (1980) 등이 제안한 검정법에 이상치 탐지법이 추가 적용된 이분산 검정과정을 제시한다. 제시된 방법은 예제 및 모의실험을 통해 기존의 이분산 검정법과 검정력을 비교한다. 2장에서는 본 연구에서 활용되는 이분산 검정들을 소개하며 3장에서는 본 연구에서 제안하는 이상치를 고려한 이분산 검정법의 절차를 설명한다. 4장에서는 예제와 모의실험의 결과가 제시되며 5장에서는 연구 결과를 요약한다.

2. 이분산 검정법

이분산 문제를 해결하기 위하여 제안된 다양한 검정방법 중 본 연구에서 활용되는 Breusch와 Pagan (1979), Goldfeld와 Quandt (1965)들이 제안한 방법들을 소개한다. 선형회귀모형에서 반응변수 Y_i와 설명변수 X_i 사이의 관계식은 다음과 같이 표현된다.

$$Y_i = X_i\beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

여기서 β는 $p \times 1$ 회귀계수벡터이고 ε_i는 평균이 0, 분산이 σ_i^2인 오차항이다.

Goldfeld와 Quandt (1965)가 제안한 이분산 검정은 x_k를 p개의 설명변수 중 특정 한 변수라고 할 때 이분산의 유형을 $\sigma_i^2 = \sigma^2 x_{ki}^2$ 형태로 가정한다. 이와 같은 유형의 이분산이 성립되면 해당 설명변수 x_k의 절대값이 클 경우 분산도 커지게 되어 설명변수의 크기로 나열된 데이터를 두 그룹으로 나눈 후 각 그룹에서 계산된 분산 추정치의 차이 유무로 이분산을 판단할 수 있다. 이에 따른 Goldfeld와 Quandt 검정 절차는 다음과 같다.

(1) 관찰치를 이분산에 관련있다고 의심되는 설명변수의 크기순으로 재 배열한다.

(2) 검정력을 높이기 위해 중앙값에 해당하는 c개의 관찰치를 제거한다.

(3) 나머지 관찰치를 각각 $(n-c)/2$개의 두 그룹으로 나눈다.

(4) 각 그룹별로 회귀모형을 추정한후 잔차제곱합(residual sum of squares; RSS)을 계산한다.

(5) 잔차제곱합으로부터 다음과 같은 검정통계량 R을 계산하고 이를 자유도 $((n-c-2p)/2, (n-c-2p)/2)$의 F분포와 비교한다.

$$R = \frac{\mathrm{RSS}_2/\mathrm{df}_2}{\mathrm{RSS}_1/\mathrm{df}_1} = \frac{\mathrm{RSS}_2}{\mathrm{RSS}_1}.$$

Breusch와 Pagan (1979)가 제안한 검정은 이분산의 유형을 보다 더 자유롭게 설정하여 Goldfeld와 Quandt (1965) 검정에서 요구되는 이분산 값의 순서화가 가능하지 않은 경우에도 적용될 수 있다. Breusch와 Pagan 검정에서 가정하는 이분산의 유형은 $\sigma_i^2 = f(Z_i\theta')$, $i = 1, \ldots n$이며, 여기서 $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{m-1})$, $Z_i = (1, z_{1i}, z_{2i}, \ldots, z_{(m-1)i})$이고 $z_{1i}, z_{2i}, \ldots, z_{(m-1)i}$은 설명변수 $x_1, x_2, \ldots, x_p$ 중 일부 이거나 전부일 수 있다. 즉 오차 분산은 절편을 포함한 비확률 변수 $z_1, z_2, \ldots, z_{m-1}$의 일차선

형식과 임의의 함수 관계에 있다. 이와 같은 이분산 유형에서 이분산 여부에 대한 가설은 다음과 같이 설정된다.

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_{m-1} = 0,$$
$$H_1 : \text{적어도 한 개 이상의 } \theta_i\text{는 0이 아니다.}$$

Breusch와 Pagan (1979) 검정의 절차는 다음과 같다.

(1) 선형모형 $y = X\beta + \varepsilon$을 추정한 후 오차에 대한 추정치인 잔차 $\hat{\varepsilon}_i$ 를 계산한다.

(2) 잔차제곱과 잔차제곱평균 $\tilde{\hat{\varepsilon}}^2$을 이용하여 통계량 $\tau_i = \hat{\varepsilon}_i^2/\tilde{\hat{\varepsilon}}_i^2$을 계산한다.

(3) τ_i를 z_{ti}에 대하여 선형회귀 적합 시킨 후 모형 설명제곱합(explained sum of squares; ESS) $= \sum_i^n (\hat{\tau}_i - \bar{\tau})^2$을 계산한다.

(4) 설명제곱합으로부터 검정통계량 $\lambda = \text{ESS}/2$을 계산하고 이를 자유도 $(m-1)$인 카이제곱 분포와 비교하여 가설검정을 수행한다.

Breusch와 Pagan 검정의 이분산의 유형중 특정한 한 유형을 가정한 검정이 White (1980) 검정이다. White 검정은 이분산이 1을 포함한 모든 x_k와 x_k^2 그리고 x_k, x_l의 교차항 등의 선형함수, 즉 $\sigma_i^2 = \beta_0 + \sum_k \beta_k x_{ki} + \sum_k \gamma_k x_{ki}^2 + \sum_{k\neq l} \delta_{kl} x_{ki} x_{li}$라고 가정한다. White 검정에서는 Breusch와 Pagan 검정에서 사용한 $\tau_i = \hat{\varepsilon}_i^2/\tilde{\hat{\varepsilon}}_i^2$ 대신 $\hat{\varepsilon}_i^2$을 모든 x_k, x_k^2, $x_k x_l$에 대해 선형 적합 시킨 후 추정된 모형의 결정계수 R^2에 표본크기를 곱한 통계량 $\psi = n \times R^2$을 $\chi^2_{(p+p(p+1)/2)}$ 분포와 비교하여 이분산에 대한 가설 검정을 수행한다. White 검정은 Breusch와 Pagan 검정과 달리 오차의 정규성 가정이 필요하지 않다.

3. 강건 이분산 검정법

모형추정이나 가설검정등 통계분석에서 이상치에 대응하는 접근법은 크게 두 가지로 나누어 진다. 첫 번째는 강건한 모수 추정치나 검정통계량을 사용하는 것이다. 예를 들면 Goldfeld와 Quandt 검정에서 관찰치를 나열하는 과정과 구분된 두 개의 그룹에 적용하는 회귀모형 추정 과정에서 비강건 요소를 강건 요소로 대체할 수 있다. 두 번째는 이상치를 탐지하여 이를 제거하는 것이다. Rana 등 (2008)은 Goldfeld와 Quandt 검정에서 이상치 탐지 접근법을 적용하여 수정된 검정절차를 제안하였다. Goldfeld와 Quandt 검정에서 분리된 두 집단의 회귀모형을 추정할 때 최소절사제곱(least trimmed of square; LTS) 추정법을 수행하여 일정한 기준이상을 초과하는 잔차의 관찰치를 이상치로 판단하여 이를 제거한다. 이상치가 제거된 관찰치들로써 OLS를 수행하며 원래의 Goldfeld와 Quandt 검정과 달리 각 그룹에서 계산된 제거잔차제곱의 중위수(median of the squared deletion residuals; MSDR)들의 비를 검정통계량으로 사용한다. 이 방법은 LTS 회귀추정 후 이상치를 판정하는 기준에 따라 검정의 강건성이 영향을 받고 검정력이 일정하지 않는 결과가 확인된다.

본 연구에서는 Goldfeld와 Quandt 검정과 Breusch와 Pagan 검정이 선형회귀모형을 기반으로 수행되므로 이상치를 탐지, 제거하는 방법으로 선형회귀모형에서 사용되는 순차적 탐지법을 제안한다. 선형회귀모형 추정에서 이상치 탐지를 위한 다양한 방법들은 이상치의 크기를 사전에 정해두는 방법과 순차적 검정을 통해 이상치의 크기를 결정하는 방법으로 구분할 수 있으며 전자에 속하는 방법들은 Gentleman과 Wilk (1975), Marasinghe (1985), Paul과 Fung (1991) 등이 있고 후자에 속하는 방법들은 Hadi와 Simonoff (1993), Kianifard와 Swallow (1989, 1990) 등이 있다. 순차적 방법은 각 단계에서 임의의 이상치군을 기반으로 새로운 이상치군을 탐지하게 되는데, 그 이전 단계에서 탐지한 이상치군을 임의의 이상치군으로 사용하는 방법과 이와 상관 없이 독립적으로 이상치군을 탐지하는 방법으로 구분된

다. 여러 방법들에 대한 장단점은 Jajo (2005) 등에 의해 설명, 비교 되었으며 본 연구에서는 다양한 방법중 계산량과 정확성의 관점에서 효율적이라고 평가되는 Hadi와 Simonoff (1993)가 제안한 단계적 탐지법을 사용한다.

본 연구에서 제안하는 강건 이분산 검정은 기존의 Goldfeld와 Quandt 검정과 Breusch와 Pagan 검정에 순차적 이상치군 탐지법을 적용하여 이상치를 제거한 후 검정을 수행하는 것이다. 순차적 이상치군 탐지법은 기초양호군으로부터 시작하여 이상치군의 크기를 한 개씩 줄여가면서 순서통계량에 기초한 t 검정을 적용하여 이상치 여부를 결정한다. 순차적 이상치군 탐지법을 적용한 수정된 Goldfeld와 Quandt 검정 과정은 다음과 같다.

(1) 관찰치를 이분산에 관련있다고 의심되는 설명변수의 크기순으로 재 배열한다.

(2) 검정력을 높이기 위해 중앙값에 해당하는 c개의 관찰치를 제거한다.

(3) 나머지 관찰치를 크기 기준으로 각각 $(n-c)/2$개의 두그룹으로 나눈다.

(4) 각 그룹별로 다음과 같은 절차에 따라 이상치를 탐지한다.

- 일정한 기준에 따라 크기 s개의 양호치군을 생성하고 이들을 M이라고 하자.
- 양호치 군에 의해 회귀모형을 추정하고 양호치군과 이상치군에 따라 각 관찰치의 내적 스튜던트화 잔차(internally studentized residual) d_i를 계산한다.

$$d_i = \begin{cases} \dfrac{y_i - x_i^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 - x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \in M, \\ \dfrac{y_i - x_i^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 + x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \notin M. \end{cases}$$

- $|d_i|$의 오름차순 j번째 순서 통계량인 $|d|_{(j)}$에 대하여 만약 $|d|_{(s+1)} \geq t_{(\alpha/\{2(s+1)\},s-p)}$이면 $(n-s)$개의 후 순위 순서통계량의 관찰치를 최종적인 이상치군으로 판단하고 그렇지 않으면 양호치군의 숫자를 한 개 늘려서 앞의 절차를 반복한다.

(5) 각 그룹별로 최종적으로 이상치라고 판정된 관찰치군을 제거한 후 회귀모형을 추정하고 Goldfeld와 Quandt 검정 통계량에 의하여 F 검정을 수행한다.

순차적 이상치 탐지방법에서 초기 양호치 선정은 기존에 제안된 잔차를 이용한 여러 방법 (Hadi와 Simonoff, 1993)을 적용하여 선정할 수 있다.

Breusch와 Pagan 검정에서도 강건성을 보완하기 위하여 Goldfeld와 Quandt 검정에서와 유사하게 이상치 탐지과정을 추가한다. Breusch와 Pagan 검정에서 τ_i를 z_{ti}에 대하여 모형을 적합하는 단계에서 순차적 이상치군 탐지법을 적용한다. 이상치가 탐지되면 이상치를 제거한 후 원래의 Breusch와 Pagan 검정 과정을 수행한다. τ_i를 z_{ti}에 대하여 모형 적합하는 단계에서 이상치를 탐지하는 것은 분산유형의 가정에서 벗어나는 것을 이상치로 간주하는 것을 의미한다. 이상치를 탐지한 후 Breusch와 Pagan 검정의 첫 번째 단계부터 새롭게 시작할 수도 있으나 이러한 절차의 결과는 앞서 설명한 과정보다 검정력이 낮은 것으로 나타났다.

4. 예제 및 모의실험

본 연구에서 제안한 방법의 검정력을 확인하기 위하여 네 개의 예제와 모의실험을 수행한다. 예제에 사

Table 4.1. Original and modified housing expenditures data

Index	Income	Hosing EXP.	Index	Income	Hosing EXP.
1	5	1.8(4.9)	11	15	4.2
2	5	2.0	12	15	4.2
3	5	2.0	13	15	4.5
4	5	2.0	14	15	4.8
5	5	2.1	15	15	5.0
6	10	3.1	16	20	4.8
7	10	3.2	17	20	5.0
8	10	3.5	18	20	5.7
9	10	3.5	19	20	6.0
10	10	3.6	20	20	6.2(2.0)

Table 4.2. P-values of heteroscedasticity test based on original Goldfeld-Quandt test (GQ-O), robust Goldfeld-Quandt (GQ-S), original Breusch-Pagan test (BP-O) and robust Breusch-Pagan (BP-S) for housing expenditure data with and without outliers

Outliers	GQ-O	GQ-S	BP-O	BP-S
Without	0.0032	0.0032	0.0079	0.0001
With	0.3168	0.0080	0.6371	0.0000

용되는 자료들은 기존의 여러 연구에서 이분산의 존재가 확인되었으며 원자료와 원자료를 수정하여 임의의 이상치를 생성한 자료에 대하여 이분산검정을 수행하여 각 검정의 검정력을 확인하기로 한다.

첫 번째 예제의 자료는 Pindyck와 Rubinfeld (1998)의 주거비 자료이다. $n = 20$개의 서로 다른 임금 집단의 주거비용에 대한 원 자료와 임의로 이를 수정하여 이상치를 생성한 자료는 Table 4.1과 같다. Table 4.1에서 괄호안의 숫자는 이상치에 해당하는 관찰치이다. 주거비자료에 대한 Goldfeld와 Quandt 검정(GQ-O)과 Breusch와 Pagan 검정(BP-O) 그리고 이상치를 고려한 수정된 Goldfeld와 Quandt 검정(GQ-S)과 수정된 Breusch와 Pagan 검정(BP-S) 결과 계산된 유의확률값은 Table 4.2와 같다. 이상치가 없을 때 네 개의 모든 검정이 이분산을 탐지하지만 이상치가 존재할 때 기존의 Goldfeld와 Quandt 검정과 Breusch와 Pagan 검정이 이분산을 탐지하는것에 실패하는데 반해 본 연구에서 제안한 수정 검정들은 성공적으로 이분산을 탐지한다.

두 번째 예제의 자료는 소비 지출비 자료 (Gujarati, 2002)이며 각 개인 임금과 소비 지출비의 $n = 30$개로 구성된다. 소비지출비 원 자료와 이상치 생성을 위해 세 개의 관찰치를 임의로 수정한 자료는 Table 4.3과 같다. 소비 지출비 자료에 대한 이분산 검정결과는 Table 4.4와 같다. 첫 번째 예제와 같이 이상치가 없을 때 네 개의 모든 검정이 이분산을 탐지하지만 이상치가 존재할 때 본 연구에서 제안한 수정 검정들만 이분산을 제대로 탐지한다.

Table 4.5에 표시된 세 번째 예제의 자료는 $n = 30$개의 광고홍보비에 따른 매출액에 관한 음식점 매출자료 (Montgomery 등, 2001)이다.

음식점 매출자료에 대한 이분산 검정 결과인 Table 4.6에서도 수정된 검정들의 이상치에 대한 강건성을 확인 할 수 있다.

네 번째 예제에 사용할 자료 (Table 4.7)는 이분산이 확인된 한 나라의 31년 동안 수입과 저축에 대한 자료 (Koutsoyiannis, 2001)이다. 수입과 저축 자료에 대한 이분산 검정결과 의도적으로 이상치를 발생시킨 자료에 대하여 수정된 검정은 이상치에 강건한 결과를 보여준다 (Table 4.8).

Table 4.3. Original and modified consumption expenditure data

Index	Expenditure	Income	Index	Expenditure	Income	Index	Expenditure	Income
1	55(10)	80	11	74	105	21	152	220
2	65(10)	100	12	110	160	22	144	210
3	70	85	13	113	150	23	175	245
4	80	110	14	125	165	24	180	260
5	79	120	15	108	145	25	135	190
6	84	115	16	115	180	26	140	205
7	98	130	17	140	225	27	178	265
8	95	140	18	120	200	28	191	270
9	90	125	19	145	240	29	137	230
10	75	90	20	130	185	30	189(100)	250

Table 4.4. P-values of heteroscedasticity test based on original Goldfeld-Quandt test (GQ-O), robust Goldfeld-Quandt (GQ-S), original Breusch-Pagan test (BP-O) and robust Breusch-Pagan (BP-S) for consumption expenditure data with and without outliers

Outliers	GQ-O	GQ-S	BP-O	BP-S
Without	0.0418	0.0418	0.0452	0.0001
With	0.6121	0.0258	0.7506	0.0000

Table 4.5. Original and modified restaurant food sales data

Index	Income	Ad. Exp.	Index	Income	Ad. Exp.	Index	Income	Ad. Exp.
1	81464(300000)	3000	11	131434	9000	21	178187	15050
2	72661	3150	12	140564	11345	22	185304	15200
3	72344	3085	13	151352	12275	23	155931	15150
4	90743	5225	14	146926	12400	24	172579	16800
5	98588	5350	15	130963	12525	25	188851	16500
6	96507	6090	16	146630	12310	26	192424	17830
7	126574	8925	17	147041	13700	27	203112(300000)	19500
8	114133	9015	18	179021	15000	28	192482	19200
9	115814	8885	19	166200	15175	29	218715	19000
10	123181	8950	20	180732	14995	30	214317(21431)	19350

Table 4.6. P-values of heteroscedasticity test based on original Goldfeld-Quandt test (GQ-O), robust Goldfeld-Quandt (GQ-S), original Breusch-Pagan test (BP-O) and robust Breusch-Pagan (BP-S) for restaurant food sales data with and without outliers

Outliers	GQ-O	GQ-S	BP-O	BP-S
Without	0.0624	0.0624	0.0368	0.0077
With	0.4604	0.0554	0.7563	0.0000

이상치가 존재할 때 제시된 모든 예제에서 수정검정들은 성공적으로 이분산을 탐지한다. 이상치가 없을 때 수정된 Goldfeld와 Quandt 검정은 어떤 관찰치도 이상치로 탐지하지 않아 원래의 Goldfeld와 Quandt 검정과 동일한 결과를 보여주지만 수정된 Breusch와 Pagan 검정은 몇 개의 관찰치를 이상치로 탐지한다. 유의확률을 기준으로 판단할 때 수정된 Breusch와 Pagan 검정이 수정된 Goldfeld와 Quandt 검정보다 검정력이 더 높은것을 알수 있다. 예제처럼 설명변수의 값들이 가장 작은 값이나 큰 값들에서 이상치를 설정된 경우와 달리 이상치가 설명변수의 중앙에서 발생하는 경우에는 각 방법들간에 유의한 차이가 발생하지 않았다.

Table 4.7. Personal saving and income data (£m)

Period	Saving(y)	Income(x)	Period	Saving(y)	Income(x)
1	264(2644)	8777	17	1578	24127
2	105(1050)	9210	18	1654	25604
3	90	9954	19	1400	26500
4	131	10508	20	1829	27670
5	122	10979	21	2200	28300
6	109	11912	22	2017	27430
7	406	12747	23	2105	29560
8	503	13499	24	1600	28150
9	431	14269	25	2250	32100
10	588	15522	26	2420	32500
11	898	16730	27	2570	35250
12	950	17663	28	1720	33500
13	779	18575	29	1900	36000
14	819	19635	30	2100(2.100)	36200
15	1222	21163	31	2300(2.300)	38200
16	1702	22880			

Table 4.8. P-values of heteroscedasticity test based on original Goldfeld-Quandt test (GQ-O), robust Goldfeld-Quandt (GQ-S), original Breusch-Pagan test (BP-O) and robust Breusch-Pagan (BP-S) for personal saving and income data with and without outliers

Outliers	GQ-O	GQ-S	BP-O	BP-S
Without	0.0066	0.0066	0.0011	0.0000
With	0.4016	0.0025	0.3567	0.0000

제안된 검정의 이상치에 대한 강건성을 검증하기 위하여 모의실험을 수행한다. 모의실험에 사용될 자료는 다음과 같은 모형으로부터 임의로 생성한다.

$$Y = X + \varepsilon,$$

오차항은 $\varepsilon \sim N(0, x^2)$에 의하여 생성되어 이분산을 갖게 되며 독립변수 X는 비확률적 추출과 확률적 추출 두 가지 방식으로 생성되고 1과 10 사이의 정수로 반복 구성되는 것과 t 분포에서 임의로 추출하는 것을 각각 사용하기로 한다. 생성되는 표본의 크기는 $n = 20, 30, 50, 80, 100$인 다섯 경우를 고려하고 표본에서 차지하는 이상치의 비율은 5%, 10%로 하며 이상치는 추세를 벗어난 관찰치로 정의하여 $Y = X - 10 + \varepsilon$, $\varepsilon \sim N(0, 0.05^2)$으로부터 생성한다.

Table 4.9와 Table 4.10은 독립변수의 값이 균등하게 생성된 경우와 t 분포로부터 생성된 경우의 모의실험에서 나타난 각 검정들의 검정력을 보여준다. 이상치가 존재할 때 기존의 Goldfeld와 Quandt 검정, Breusch와 Pagan 검정이 매우 낮은 검정력을 보이는 것과는 달리 수정된 검정들은 이보다 훨씬 개선된 결과를 보인다 (Table 4.9). 특히 이 경우 이상치 탐지법의 추가는 Goldfeld와 Quandt 검정보다 Breusch와 Pagan 검정이 더 효과적임을 알 수 있다. Table 4.10도 유사한 결과를 보여준다.

5. 결론

이상치를 제거하여 이분산 검정을 수행하는 접근법은 다양하게 시도될 수 있다. Rana 등 (2008)이 제시한 방법도 본 연구와 유사하게 예제와 모의시험을 수행하였으나 대체로 본 연구에서 제시한 방법보다

Table 4.9. Simulation results of heteroscedasticity tests for X being taken equally spaced

% of outliers	n	GQ-O	GQ-S	BP-O	BP-S
5	20	0.0352	0.3136	0.0866	0.8213
	30	0.0335	0.2982	0.0770	0.9252
	50	0.0281	0.3251	0.0495	0.9691
	80	0.0163	0.3227	0.0521	0.9923
	100	0.0206	0.4048	0.0457	0.9965
10	20	0.0175	0.2462	0.0462	0.7863
	30	0.0258	0.2765	0.0345	0.8804
	50	0.0172	0.3157	0.0521	0.9574
	80	0.0133	0.3013	0.0472	0.9908
	100	0.0116	0.3036	0.0496	0.9967

GQ-O = Goldfeld-Quandt test, GQ-S = robust Goldfeld-Quandt, BP-O = original Breusch-Pagan test, BP-S = robust Breusch-Pagan.

Table 4.10. Simulation results of heteroscedasticity tests for X being taken from t-distribution

% of outliers	n	GQ-O	GQ-S	BP-O	BP-S
5	20	0.2432	0.3785	0.4335	0.9750
	30	0.4435	0.5097	0.5123	0.9948
	50	0.1266	0.5663	0.7397	0.9998
	80	0.0422	0.6587	0.7245	1.0000
	100	0.0748	0.7339	0.8518	1.0000
10	20	0.0226	0.3605	0.5049	0.9788
	30	0.1797	0.5682	0.6632	0.9973
	50	0.1117	0.6456	0.7345	0.9999
	80	0.0398	0.7558	0.7463	1.0000
	100	0.0292	0.7711	0.8788	1.0000

GQ-O = Goldfeld-Quandt test, GQ-S = robust Goldfeld-Quandt, BP-O = original Breusch-Pagan test, BP-S = robust Breusch-Pagan.

검정력이 낮은 것으로 확인된다. 모의실험 결과 Goldfeld와 Quandt 검정은 이분산 유형이 가정된 모양을 따른다 하더라도 이상치가 존재할 경우 검정력이 대단히 낮으며 Breusch와 Pagan 검정도 이상치의 유형에 따라 같은 현상을 보인다. 자료의 분포와 이상치의 분포에 따라 차이가 있지만 어떤 경우에도 본 연구에서 제안된 검정절차는 이상치에 강건함을 알 수 있다. 본 연구에서는 이상치를 추세 이탈 관찰치로 정의하였으나 이분산의 유형이 다른 관찰치를 이상치로 간주하여 유사한 절차를 개발 할 수 있으며 본 연구에서 사용한 접근법은 Goldfeld와 Quandt 검정과 Breusch와 Pagan 검정 이외 어떠한 이분산 검정에도 적용할 수 있다.

References

Anscombe, F. J. (1961). Examination of residuals, In *Proceedings of 4*$^{\text{th}}$ *Berkeley Symposium*, **1**, 1–36.

Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Oxford University Press, Oxford.

Bickel, P. (1978). Using residuals robustly I: tests for heteroscedasticity, nonlinearity, *Annals of Statistics*, **6**, 266–291.

Breusch, T. and Pagan, A. (1979). A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47**, 1287–1294.

Carroll, R. J. and Ruppert, D. (1981). On robust tests for heteroscedasticity, *Annals of Statistics*, **9**, 205–209.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, 2nd ed., Wiley, New York.

Cheng, T.-C. (2012). On simultaneously identifying outliers and heteroscedasticity without specific form, *Computational Statistics and Data Analysis*, **56**, 2258–2272.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression, *Biometrika*, **70**, 1–10.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed., John Wiley, New York.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers II: supplementing the direct analysis of residuals, *Biometrics*, **31**, 387–410.

Giloni, A., Simonoff, J. S., and Sengupta, B. (2006). Robust weighted LAD regression, *Computational Statistics and Data Analysis*, **50**, 3124–3140.

Goldfeld, S. M. and Quandt, R. E. (1965). Some tests for homoscedasticity, *Journal of the American Statistical Association*, **60**, 539–547.

Gujarati, D. (2002). *Basic Econometrics*, 4th ed., McGraw-Hill, New York.

Hadi, A. S., and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hammerstrom, T. (1981). Asymptotically optimal tests for heteroscedasticity in the general linear model, *Annals of Statistics*, **9**, 368–380.

Horn, P. (1981). Heteroscedasticity of residuals: a non-parametric alternative to the Goldfeld-Quandt peak test, *Communications in Statistics-Theory and Methods*, **10**, 795–808.

Hubert, M. and Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference*, **57**, 153–163.

Jajo, N. K. (2005). A review of robust regression an diagnostic procedures in linear regression, *Acta Mathematicae Applicatae Sinica*, **21**, 209–224.

Kianifard, F. and Swallow, W. H. (1989). Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression, *Biometrics*, **45**, 571–585.

Kianifard, F. and Swallow, W. H. (1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression, *Communications in Statistics-Theory and Methods*, **19**, 1913–1938.

Koutsoyiannis, A. (2001). *Theory of Econometrics*, 2nd ed., Palgrave, New York.

Marasinghe, M. G. (1985). A multistage procedure for detecting several outliers in linear regression, *Technometrics*, **27**, 395–399.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*, 3rd ed., Wiley, New York.

Paul, S. R. and Fung, K. Y. (1991). A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression, *Technometrics*, **33**, 339–348.

Pindyck, S. R. and Rubinfeld, L. D. (1998). *Econometric Models and Econometric Forecasts*, 4th ed., Irwin/McGraw-Hill, New York.

Rana, M. S., Midi, H., and Imon, A. H. M. R. (2008). A robust modification of the Goldfeld-Quandt test for the detection of heteroscedasticity in the presence of outliers, *Journal of Mathematics and Statistics*, **4**, 277–283.

Ryan, T. P. (2008). *Modern Regression Methods*, 2nd ed., Wiley, New York.

Weisberg, S. (2005). *Applied Linear Regression*, Wiley, New York.

White, H. (1980). Heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity, *Econometrica*, **48**, 817–838.

이상치 탐지법을 이용한 강건 이분산 검정

서한손[a] · 윤민[b,1]

[a]건국대학교 응용통계학과, [b]부경대학교 통계학과

(2015년 11월 2일 접수, 2016년 2월 10일 수정, 2016년 3월 12일 채택)

요 약

회귀분석에서 이분산이 발생할 경우 표준적 추정절차에 따른 결과는 유효하지 않게 되므로 이를 확인하는 것이 필요하다. 이분산 문제와 더불어 이상치가 함께 존재하면 이분산에 관한 진단은 왜곡될 수 있다. 이상치가 존재할 때 이분산을 진단하는 기존의 방법들은 강건통계량을 이용하거나 이상치를 제거하는 접근법을 사용한다. 이분산 문제에서 이상치를 탐지하기 위하여 여러 가지 접근법이 제시되었다. 본 연구에서는 이분산 진단과정에서 이상치를 배제하기 위하여 기존의 이분산 검정과정에 순차적 이상치 탐지법을 적용하는 절차를 제시한다. 제시된 방법은 모의실험 및 예제를 통해 기존의 검정방법과 검정력을 비교한다.

주요용어: 강건 검정, 선형회귀모형, 이분산, 이상치

이 논문은 2015학년도 건국대학교의 연구년교원 지원에 의하여 연구되었음.

[1]교신저자: (48513) 부산시 남구 용소로 45, 부경대학교 통계학과. E-mail: myoon@pknu.ac.kr