International Journal of Statistics and Probability; Vol. 4, No. 3; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On the Detection of Heteroscedasticity by Using CUSUM Range Distribution

A. Nanthakumar[1], S. Kanbur[2], & E. Wilson[1]

[1] Department of Mathematics, SUNY-Oswego, NY 13126, USA

[2] Department of Physics, SUNY-Oswego, NY 13126, USA

Correspondence: A. Nanthakumar, Department of Mathematics, SUNY-Oswego, NY 13126, USA. Tel: 1-315-312-2738. E-mail: ampala.nanthakumar@oswego.edu

Received: April 23, 2015 Accepted: May 20, 2015 Online Published: July 1, 2015

doi:10.5539/ijsp.v4n3p88 URL: http://dx.doi.org/10.5539/ijsp.v4n3p88

Abstract

In this paper, we present a new method for checking the heteroscedasticity among the error terms. The method is based on the CUSUM Range distribution. We derive the CUSUM Range distribution (under the assumption of homogeneity) and use it to test for heteroscedasticity. The method seems to detect heteroscedasticity when it is present among the error terms.

Keywords: heteroscedasticity, CUSUM, range

1. Introduction

The problem of detecting heteroscedasticity among the uncorrelated error terms has been an interesting topic for many years in Statistical, Econometric and Financial analysis. In the past, the hetoeroscedasticity was studied by several people. Morgan (1939), Pitman (1939), Wilks (1946) were some of the pioneers to investigate heteroscedasticity in the variability of the error terms. Later Valavanis (1959), Cacoullos (1965), Cacoullos (2001), Goldfeld & Quandt (1973), Chen & Gupta (1997), Kanbur et al. (2010), Kanbur et al. (2009) are among many others who have studied the heteroscedasticity in various contexts. In regression, the residuals can be tested for heteroscedasticity by using the tests such as Pittman-Morgan t-test or the Breusch-Pagan test which regresses the squared residuals to the independent variables. However, the Breusch-Pagan test is very sensitive to the error normality and hence a more robust Koenkar-Basset test (or the generalized Breusch-Pagan test) is preferred. For testing group-wise heteroscedasticity, the Goldfeldt-Quandt test and the Levene test are commonly used.

Kanbur et al. (2007) developed a nonparametric test for heteroscedasticity of the residual errors by using the CUSUM ranks. However, this test uses bootstrap based numerical simulations in this nonparametric approach to test for heteroscedasticity. Generally speaking, the nonparametric tests are less powerful. So, in this paper, we develop a theoretical (parametric) test based on the CUSUM Range. This test is theory based and so does not require bootstrap as a tool to detect heteroscedasticity. We use the same data that is used in Kanbur et al. (2007) for the comparison.

2. Methodology

We will analyze the error residuals of a simple linear regression in order to check for a possible heteroscedasticity among the error variability. Here, the response variable is the Cepheid Luminosity and the predictor is the natural logarithm of the Cepheid Periodicity. The Cepheids are pulsating stars from the nearby galaxy LMC (Large Magellanic Clouds).

$$L_i = \alpha + \beta . \log(P_i) + \epsilon_i \tag{1}$$

for $i = 1, 2, ..., n$ where L represents the luminosity; P the periodicity and ϵ, the noise in the data. We use the least square method to estimate the unknowns.

Let

$$C(j) = \sum_{i=1}^{j}(L_i - a - b. \log(P_i)) = \sum_{i=1}^{j} \epsilon_i \tag{2}$$

where a and b are the least square estimates of α and β respectively.

As noted in Koen et al (2007) when there are no deviations from the linearity (or in other words when the error variability is homoscedastic) then $C(j)$ is the sum of the uncorrelated random variates and hence it is a random walk. On the other hand, when there is a violation then $C(j)$ will not be a random walk. Here, we will use the range statistic

$$R = \max C(j) - \min C(j) \tag{3}$$

Next, we present a table for the possible values of the range statistic. This table is very helpful in constructing the probability distribution for the Range Statistic. Note that n represents the sample size.

Table 1. $n = 2$

Min	Max	R	Probability
ϵ_1	$\epsilon_1 + \epsilon_2$	ϵ_2	0.5
$\epsilon_1 + \epsilon_2$	ϵ_1	$-\epsilon_2$	0.5

Table 2. $n = 3$

Min	Max	R	Probability
ϵ_1	$\epsilon_1 + \epsilon_2$	ϵ_2	0.125
$\epsilon_1 + \epsilon_2$	ϵ_1	$-\epsilon_2$	0.125
$\epsilon_1 + \epsilon_2$	$\epsilon_1 + \epsilon_2 + \epsilon_3$	ϵ_3	0.125
$\epsilon_1 + \epsilon_2 + \epsilon_3$	$\epsilon_1 + \epsilon_2$	$-\epsilon_3$	0.125
ϵ_1	$\epsilon_1 + \epsilon_2 + \epsilon_3$	$\epsilon_2 + \epsilon_3$	0.250
$\epsilon_1 + \epsilon_2 + \epsilon_3$	ϵ_1	$-\epsilon_2 - \epsilon_3$	0.250

Table 3. $n = 4$

Min	Max	R	Probability
ϵ_1	$\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$	$\epsilon_2 + \epsilon_3 + \epsilon_4$	2/11
$\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$	ϵ_1	$-\epsilon_2 - \epsilon_3 - \epsilon_4$	2/11
$\epsilon_1 + \epsilon_2$	$\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$	$\epsilon_3 + \epsilon_4$	1/11
$\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$	$\epsilon_1 + \epsilon_2$	$-\epsilon_3 - \epsilon_4$	1/11
$\epsilon_1 + \epsilon_2 + \epsilon_3$	ϵ_1	$-\epsilon_2 - \epsilon_3$	1/11
ϵ_1	$\epsilon_1 + \epsilon_2 + \epsilon_3$	$\epsilon_2 + \epsilon_3$	1/11
ϵ_1	$\epsilon_1 + \epsilon_2$	ϵ_2	1/22
$\epsilon_1 + \epsilon_2$	ϵ_1	$-\epsilon_2$	1/22
$\epsilon_1 + \epsilon_2 + \epsilon_3$	$\epsilon 1 + \epsilon 2 + \epsilon_3 + \epsilon_4$	ϵ_4	1/22
$\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon 4$	$\epsilon 1 + \epsilon_2 + \epsilon_3$	$-\epsilon_4$	1/22
$\epsilon_1 + \epsilon_2$	$\epsilon_1 + \epsilon 2 + \epsilon 3$	ϵ_3	1/22
$\epsilon_1 + \epsilon_2 + \epsilon_3$	$\epsilon_1 + \epsilon_2$	$-\epsilon_3$	1/22

Remark: Based on the probability weight pattern, we have developed a formula for the probability weight

$$w_n(i) = \frac{(n-i)2^i}{(2^{n+1} - 2(n+1))} \tag{4}$$

where n equals the number of observations and i equals the number of error terms in the CUSUM Range.

Next, we derive the distribution for the Range (R). Note that the range R can be written in the following form as the error terms are exchangeable (due to being independent and identical in distribution).

$$R = \max(0, \sum_{j=1}^{i} \epsilon_j) \tag{5}$$

Given $i = 1$ then

$$P(R \leq x) = P(0 < \epsilon_1 \leq x | \epsilon_1 > 0) = 2[\phi(\frac{x}{\sigma_\epsilon}) - 0.5]. \tag{6}$$

Given $i = 2$ then

$$P(R \leq x) = P(0 < \epsilon_1 + \epsilon_2 \leq x | \epsilon_1 + \epsilon_2 > 0) = 2[\phi(\frac{x}{\sqrt{2}\sigma_\epsilon}) - 0.5]. \tag{7}$$

In the general case, given there are i number of error terms then,

$$P(R \leq x) = P(0 < \epsilon_1 + \epsilon_2 + ... + \epsilon_i \leq x | \epsilon_1 + \epsilon_2 + ... + \epsilon_i > 0) = 2[\phi(\frac{x}{\sqrt{i}\sigma_\epsilon}) - 0.5]. \tag{8}$$

Remark: Overall, the CUSUM Range is a mixture distribution.

Next, we present the CUSUM Range distribution.

Lemma 1. Let R represent the CUSUM Range. Then,

$$\begin{aligned} P(R \leq x) &= \sum 2w_n(i)[\phi(\frac{x}{\sqrt{i}\sigma_\epsilon}) - 0.5] \\ &= 2\sum_{i=1}^{n} \frac{(n-1)2^i}{(2^{n+1} - 2(n+1))}[\phi(\frac{x}{\sqrt{i}\sigma_\epsilon}) - 0.5] \\ &= \sum_{i=1}^{n} \frac{(n-i)2^i}{(2^n - (n+1))}[\phi(\frac{x}{\sqrt{i}\sigma_\epsilon}) - 0.5]. \end{aligned} \tag{9}$$

Proof. See Appendix Section. □

Remark: The CUSUM Range is a mixture of folded normal variates. Its density is given by

$$f_R(r) = \frac{1}{(2^n - (n+1))} \sum_{i=1}^{n} \frac{(n-1)2^i e^{-\frac{r^2}{2i\sigma_\epsilon^2}}}{\sqrt{2\pi} i \sigma_\epsilon}. \tag{10}$$

Also, the conditional density function is given by

$$f_R(r | r \geq y) = \frac{1}{(2^n - (n+1))} \sum_{i=1}^{n} \frac{(n-1)2^i e^{-\frac{r^2}{2i\sigma_\epsilon^2}}}{\sqrt{2\pi} i \sigma_\epsilon [1 - \phi[\frac{y}{\sqrt{i}\sigma_\epsilon}]]}. \tag{11}$$

Result 1: The conditional expected value for the CUSUM Range is given by

$$E(R | R \geq y) = \frac{\sigma_\epsilon}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{(n-1)2^{i-1}\sqrt{i} e^{-\frac{y^2}{2i\sigma_\epsilon^2}}}{(2^n - (n+1))[1 - \phi[\frac{y}{\sqrt{i}\sigma_\epsilon}]]}. \tag{12}$$

Result 2: The conditional second moment for the CUSUM Range is given by

$$E(R^2 | R \geq y) = \frac{\sigma_\epsilon y}{(2^n - (n+1))\sqrt{2\pi}} \sum_{i=1}^{n} \frac{(n-i)\sqrt{i} 2^{i-1} e^{-\frac{r^2}{2i\sigma_\epsilon^2}}}{[1 - \phi[\frac{y}{\sqrt{i}\sigma_\epsilon}]]} + \sigma_\epsilon^2 \sum_{i=1}^{n} \frac{i(n-i)2^i}{(2^n - (n+1))}. \tag{13}$$

As we noted earlier when there is homogeneity among the error variability, the Range follows a mixture of folded normal distribution. Hence its conditional expected value and the conditional second moment satisfy equations (12) and (13).

In the next section, we present the numerical results.

3. Numerical Results

Here in this section, we present the numerical results based on the simulated data and the actual data.

3.1 Simulated Data

The following table presents the numerical values (based on the simulated data) for the conditional expected values given by (12) and the simulation based empirical estimates for the conditional expected values.

Case 1: No change in error variability ($n = 10, \sigma_\epsilon = 0.022$)

A random sample of error terms of size = 10 was simulated according to a normal distribution with mean = 0 and standard deviation $\sigma_\epsilon = 0.022$. This sample of size = 10 was simulated several times to compute the empirical estimate.

Table 4.

r	Conditional Expected Value (Formula)	Conditional Expected Value (Empirical)
0.05	0.082163	0.088957
0.06	0.090130	0.094583
0.07	0.098293	0.098213
0.08	0.106630	0.102915
0.09	0.115120	0.107997
0.10	0.123740	0.117707
0.11	0.132490	0.130043

From the simulation, it is pretty clear that there is no change in the error variability as is the case with this simulation. The formula based conditional expected values are fairly close to the empirical estimates.

Case 2: Change in error variability ($n = 20, \sigma_{1\epsilon} = 0.022, \sigma_{2\epsilon} = 0.045$)

A random sample of 20 error terms was simulated several times with 5 of these error terms following a normal distribution with mean = 0 and standard deviation $\sigma_{1\epsilon} = 0.022$, and the other 15 error terms following a normal distribution with mean = 0 and standard deviation $\sigma_{2\epsilon} = 0.045$.

Table 5.

r	Conditional Expected Value (Formula)	Conditional Expected Value (Empirical)
0.10	0.1872	0.14574
0.11	0.1888	0.15393
0.12	0.2019	0.16223
0.13	0.2059	0.17063
0.14	0.2098	0.17913
0.15	0.2224	0.18772
0.16	0.2358	0.19639
0.17	0.2472	0.20514
0.18	0.2502	0.21396
0.19	0.2567	0.22284
0.20	0.2775	0.23178
0.21	0.2692	0.24079
0.22	0.2968	0.24984
0.23	0.3220	0.25895
0.24	0.3349	0.26810
0.25	0.3756	0.27729

As we can see from the simulation results, there is a difference between the formula based conditional expected values and the empirical estimates based on the simulation. This result supports the fact that there is a change in the error variability (as is the case with this simulation).

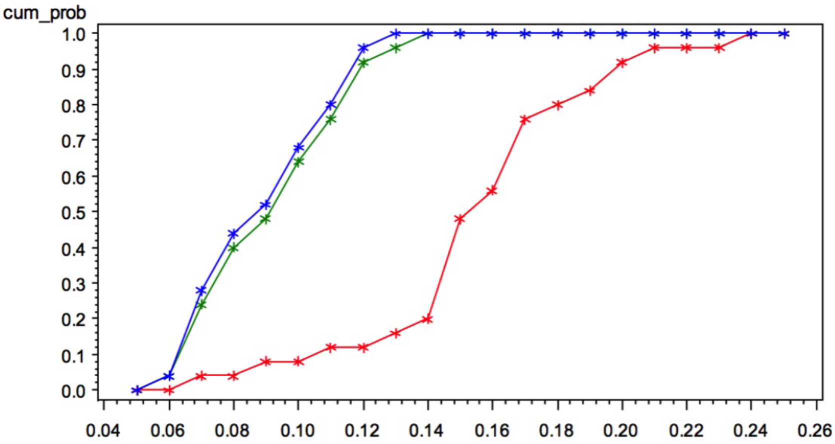

Figure 1. Graphs of cumulative distribution.

Note: The graph in blue represents the cumulative probability distribution based on the CUSUM Range from the simulated error terms. The graph in green is for the residual errors with homogeneous variance and the graph in red is for the residual errors with heterogeneous variance.

3.2 A Real Application

We can use the concept of error homogeneity (heterogeneity) to check whether the relationship between two quantitative variables is linear or not. For example, when the relationship is actually not linear (or when there is a change point) then the residual error variance will not be homogeneous under the assumption of linearity. This is the concept that we are about to use to check for a possible linear relationship between the Cepheid Period and the Luminosity in the next example.

Example (based on Actual Data)

Here, we consider an astrophysical data set that has 1779 observations about the Cepheid Period and Luminosity Relationship. In order to check whether this relationship is linear or not, we drew several random samples of size n=10 and again several random samples of size n=20 from this data set. Just like we did in the previous simulation, we computed the empirical (cumulative) distribution for the CUSUM Range and the theoretical (cumulative) distribution for the CUSUM Range based on at selected values of x as indicated below.

Table 6. For $n = 10$

x	0.3	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
Empirical	0.05	0.4	0.75	0.75	0.85	0.85	0.85	0.90	1.0
Theoretical	0.404	0.618	0.773	0.876	0.937	0.971	0.987	0.995	0.998

Table 7. For $n = 20$

x	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2
Empirical	0.1	0.35	0.65	0.70	0.90	0.90	0.90	0.95	1.0
Theoretical	0.494	0.624	0.731	0.815	0.878	0.922	0.952	0.972	0.984

The Goodness-of-Fit test and the Kolmogorov-Smirnov test clearly indicate a difference between the empirical distribution and the theoretical distribution (which assumes a linear relationship between the predictor and the response variable). This means that in this astrophysical data, the relationship is not linear.

4. Discussion and Conclusion

This paper gives a mathematical justification as to how the CUSUM Range distribution is affected by the residual error heteroscedasticity. Unlike the papers that were published in the past on the topic of error heteroscedasticity, this paper presents a very simple method to detect error heteroscedasticity. The paper uses only the assumption of normality among the error terms, and this is always an assumption in the Linear Models and to a larger extent in the Time series Models. Moreover, this research confirms as in the previous papers that the Cepheid Period-Luminosity relationship is not linear.

References

Cacoullos, T. (1965). A relation between t and F-distributions. *Journal of American Statistical Association, 60,* 528 - 531. http://dx.doi.org/10.2307/2282687

Cacoullos, T. (2001). The F-test of homoscedasticity for correlated normal variables. *Statistics & Probability Letters, 54,* 1 - 3. http://dx.doi.org/10.1016/S0167-7152(00)00189-9

Chen, J., & Gupta, A. K. (1997). Testing and locating variance change points with applications to stock prices. *Journal of American Statistical Association, 92,* 739 - 747. http://dx.doi.org/10.1080/01621459.1997.10474026

Goldfeld, S. M., & Quandt, R. E. (1973). The estimation of structural shifts by switching regressions. *Annals of Economics and Social Measurement, 2*, 475 - 485.

Kanbur, S., Koen, C., & Ngeow, C. (2007). The detailed forms of the LMC Cepheid PL and PLC relations. *Monthly Notices of the Royal Astronomical Society, 380,* 1440 - 1448. http://dx.doi.org/10.1111/j.1365-2966.2007.12101.x

Kanbur, S., Nanthakumar, A., & Ngeow, C. (2009). On the near equivalence of Testimation and Schwarz Information Criterion (SIC) to study Cepheid-Period Luminosity Relation. *Journal of Statistical Theory and Practice, 3,* 805 - 815. http://dx.doi.org/10.1080/15598608.2009.10411961

Kanbur, S., Nanthakumar, A., Ngeow, C., & Marsh, A. (2010). A Comparison of Testimation and Schwarz Information Criterion for heteroscedasticity. *Journal of Statistics and Applications, 5,* 241 - 258.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika, 31,* 13 - 19. http://dx.doi.org/10.2307/2334972

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika, 31*, 9 - 12. http://dx.doi.org/10.2307/2334971 Valavanis, D. (1959). Econometrics. McGraw-Hill.

Wilks, S. S. (1946). Sample Criteria for testing equality of means, equality of variances, equality of covariances in a multivariate distribution. *Annals of Mathematical Statistics, 17*, 257 - 281. http://dx.doi.org/10.1214/aoms/1177730940

Appendix: Proof of Lemma 1

Note that one can write,

$$P(R \leq x) = \sum_{i=1}^{n} P(R \leq x|N = i)P(N = i).$$

But,

$$P(N = i) = w_n(i) \text{ and } P(R \leq x|N = i) = 2[\phi(\frac{x}{\sqrt{i}\sigma_\epsilon}) - 0.5].$$

where $w_n(i)$ is as described in the main body of this paper and N is the number of error terms that constitute the CUSUM range.

Copyrights