ISSN(E):2522-2260
ISSN(P):2522-2252

Journal DOI: https://doi.org/10.29145/jqm

Indexing/Abstracting

Published by
Department of Quantitative Methods
School of Business and Economics
University of Management and Technology, Lahore, Pakistan

Journal of Quantitative Methods

Evaluation of Test Statistics for Detection of Outliers and Shifts

Author(s)
Amena Urooj[1], Zahid Asghar[2]

Affiliations
[1]Pakistan Institute of Development Economics, Islamabad, Pakistan
[2]Quaid e Azam University, Islamabad, Pakistan
Email: amna@pide.org.pk
ORCID: https://orcid.org/0000-0003-4626-5525

Manuscript Information

Submission Date: May 5, 2019

Publication Date: August 31, 2020

Conflict of Interest: None

Supplementary Material: No supplementary material is associated with the article

Funding: This research received no external funding

Acknowledgment: I, Amena Urooj, hereby declare that this paper is a part of my thesis.

Citation in APA Style: Urooj, A. & Asghar. Z. (2020). Evaluation of test statistics for detection of outliers and shifts, *Journal of Quantitative Methods, 4*(2), 54-75.

This manuscript contains references to 11 other manuscripts.

The online version of this manuscript can be found at
https://ojs.umt.edu.pk/index.php/jqm/article/view/278

DOI: https://doi.org/10.29145/2020/jqm/040203

Additional Information

Subscriptions and email alerts: editorasst.jqm@umt.edu.pk
For further information, please visit https://ojs.umt.edu.pk/index.php/jqm

Journal of Quantitative Methods
4(2) 54-75
https://doi.org/10.29145/2020/jqm/040203

Evaluation of Test Statistics for Detection of Outliers and Shifts

Amena Urooj[1], Zahid Asghar[2]
[1]Pakistan Institute of Development Economics, Islamabad, Pakistan
[2]Quaid-e-Azam University, Islamabad, Pakistan
Email: amna@pide.org.pk

Received: May 05, 2019, Last Revised: Aug 12, 2020 , Accepted: Aug 29, 2020

Abstract

The existence of outliers and structural breaks in time series data offer challenges to data analysts in model identification, estimation and validation. Detection of outliers of a different nature and structure is the focus of the current study. To analyze the impact of structural breaks and outliers on model identification, estimation and their inferential analysis, we use two data generating processes; MA (1) and ARMA (1, 1). The performance of the test statistics for detecting additive outlier (AO), innovative outlier (IO), level shift (LS) and transient change (TC) is investigated using simulations. For evaluation, power of test, empirical level of significance, empirical critical values, misspecification frequencies, and sampling distribution of estimators for the two models are calculated. The empirical critical values are found higher than the theoretical cut-off (C); empirical power of the test statistics is not satisfactory for small sample size, large C and large model coefficients. The confusion between LS, AO, TC, and IO assuming different C and sample sizes is also explored. Further, empirical evidence is noticed that for Pakistan using 3-stage iterative procedure to detect multiple outliers and structural breaks. It is found that neglecting shocks lead to wrong identification, biased estimation, and excess kurtosis.

Keywords: discordant observations, structural breaks, simulation analysis, additive outlier, innovative outlier, transient change, level shift, iterative procedure

JEL Classification Codes: C15, C18, C63, C32, C87, C51, C52, C82

***AMS Classification Codes*:** 62, 65, 91, DI, 62-08, 62J20, 00A72, 91-08, 91-10, 91-11 62P20, 91B82, 91B84, 62M07, 62M09, 62M10, 62M15, 62M20

[1] **Author's Acknowledgment:** I hereby declare that this paper is a part of my thesis.

1. Introduction

Time series variables are extensively used to study aggregate fluctuations in the characteristics of any phenomenon. Occurrence of sudden events causes short and long term changes in the behavior of the phenomena under study. As the knowledge about the causes of aggregate fluctuations is in interest of policy makers, the occurrence of outliers or discordant observations[2] and structural breaks is also of great interest. Outliers and structural break detection, and their impact on modeling time series data have been investigated extensively in the literature. Several methods to handle the issue of outliers have been devised based on diagnostic, robust and Bayesian approach. The widely used diagnostic approach, initially estimates the model and its parameters using maximum likelihood estimation (MLE) method and then the residuals are analyzed to detect outliers iteratively. This procedure was initially proposed by Fox (1972).

Tsay (1986) and Chang Tiao, and Chen (1988) worked on detection and estimation of unknown outliers and structural breaks using iterative procedure. It was later modified by several contributors including Pena (1990), Tsay (1988), Balke (1993), Balke and Fomby (1994), Louni (2008), Chen and Liu (1993), Kaiser and Maravall (2001) and many others. Afterwards, Chen and Liu (1993) modified it for joint outlier detection and parameter estimation. Additive outliers (AO), innovative outliers (IO), level shift(LS) and transient change(TC) are commonly considered types of outliers. For the detection of these different types of outliers, various test statistics are widely used as suggested by Tsay (1986). However, these test statistics show varying performance in different time series structures.

Main objective of this study is to examine the widely differing properties of the test statistics for detecting outliers and structural

[2]We have used the term outliers and discordant observations interchangeably for indicating the anomalous observations.

breaks under finite sample behavior. Another objective is to analyze the behavior of time series data having structural breaks and outliers and to identify the best possible model in the presence of various types of disturbances. This is achieved by focusing on the performance of these test statistics for different choices of parameters in MA(1) and ARMA (1, 1) models through simulations. The choice of these two models is postulated on the argument that these commonly used nonlinear models provide parsimonious representation of data, make easier to spot trend and remove short term noise along with AR(1) model. The performance of these test statistics for outlier detection in AR(1) process is already evaluated by Urooj and Asghar (2017), while now we look at existence, impact and detection of various types of outliers in some nonlinear models. We evaluate the performance of test statistics in detecting the outliers in some nonlinear models via simulations. Further, empirical analysis is carried on some monthly time series of Pakistan which are expected to be more sensitive to macroeconomic, social, political and environmental uncertainty yielding high variance and more outliers. This is to assess the performance of Chen and Liu (1993) procedure in terms of incidence of misidentification, intensity of masking and swamping effect in case of Pakistan. Lack of profound relevant literature also motivates us to detect outliers in time series data for Pakistan.

This study contributes to the existing literature in several ways. Firstly, simulation study examines the sampling distribution of estimators of the nonlinear contaminated series. Secondly, the vulnerability to spurious outliers and appropriateness of the cutoff points are judged through empirical level of significance and empirical critical values. Empirical power of test analyzes sensitivity of the test statistics for outliers. Count for misspecification frequencies discovers the vulnerability to masking of outliers. Thirdly, we also study the behavior of the decaying parameter (δ) in correctly detecting TC. Lastly, the application of Chen and Liu (1993) procedure for the case of Pakistan identifies discordant observations, effects of discontinuities, and provides robust estimates of the model. Hence, better insight enables effective forecasting and policy formulation.

Section 2 explores the impact of four types of outliers on MA (1) and ARMA (1, 1) models, their autocorrelation functions (ACF), estimates and residuals. Section 3 describes simulation experiment in

detail. The patterns noted under empirical level of significance, empirical critical values, empirical power of test statistics, properties of sampling distribution in the presence of one outlier and behavior of δ in detection of TC are discussed in Section 4. Section 5 elaborates the empirical analysis of outlier and structural break detection for Pakistan. Finally, the key findings and conclusion are in section 7.

2. Impact of Outliers on MA(1) and ARMA (1, 1) Model[3]

Examining the impact of AO, TC, IO and LS on the stylized characteristics of MA(1) and ARMA(1,1) process, we initiate for the observed contaminated with outliers series as

$$z_t = y_t + A_t \quad (1)$$

where $\phi(B)\Phi(B^s)\nabla^d\nabla_s^D y_t = \theta(B)\Theta(B^s)a_t$[4] is an outlier free time series and $A_t = \omega_i v_i(B) I_t^{(T)}$, $I_t{}^{(T)}$ is an indicator variable as $I_t{}^{(T)} = 1\ at\ t = T\ and\ zero\ elsewhere,$ ω_i is the magnitude of i[th] outlier, $v_i(B)$ determines the dynamics of outliers for $i = AO, IO, LS, TC$. All other terms are defined as per usual notation (See Urooj; 2016). Using the residuals after computing MLE for the model parameters on z_t, $v_i(B)$ and ω_j are estimated based upon iterative calculation of $\eta_i = \max_t \lambda_t |\hat{\tau}_i(t_1)| > C$[5], for a possibility of type *i* outlier at t_1 with

$$\hat{\tau}_i(t) = \frac{\widehat{\omega}_i}{\rho_i \sigma_a} \ ; t = 1,2, \dots, n\ ;$$

where

$\rho_{AO}^2 = (1 + \pi_1^2 + \cdots + \pi_{n-T}^2)^{-1}$, $\rho_L^2 = (1 + \eta_1^2 + \cdots + \eta_{n-T}^2)^{-1}$, $\rho_{TC}^2 = (1 + \beta_1^2 + \cdots + \beta_{n-T}^2)^{-1}$, $\rho_{IO}^2 = 1$[6]

These test statistics are originally suggested by Tsay (1988), later used in three stage outlier detection procedure by

[3]Detail derivations are not included due to space issue.

[4]The uncontaminated MA(1) model is $y_t = a_t + \theta_1 a_{t-1}$ and ARMA (1,1) model is $y_t = \phi_1 y_{t-1} + a_t + \theta_1 a_{t-1}$

[5] C is predetermined critical value/ cut-off chosen at 3, 3.5 or 4. δ ,$0 < \delta < 1$ is a prespecified constant determining the speed of decay as proposed in Tsay (1988) and Chen et al. (1993).

[6] $\eta(B) = 1 - \eta_1 B - \eta_2 B^2 - \cdots = \frac{\pi(B)}{1-B}$, $\beta(B) = 1 - \beta_1 B - \beta_2 B^2 - \cdots = \frac{\pi(B)}{1-\delta B}$, , $\frac{1}{1-\delta B} = 1 + \delta B^1 + \delta^2 B^2 + \delta^3 B^3 + \cdots$

Chen and Liu (1993) and subsequently by Kaiser and Maravall (2001).

AO is an exogenous change occurring at point T such that$1 \leq T \leq n$. It affects the observed series at one point, say t=T. In the presence of IO, the series has an endogenous change in the noise process at some point T for $1 \leq T \leq n$, the observed series with MA(1) model is affected at T and T+1 points while for ARMA(1,1) models the occurrence of IO affects the time series up to next few lags depending upon the weights $\psi(B)$. In the presence of LS, the series has a modification in its level or mean value at time T (such that for$1 \leq T \leq n$), which lasts till the end depending upon the magnitude ω_{LS}. The TC is a special kind of level shift which dies out exponentially. In the presence of TC, the series has a temporary modification in its level starting at time T (such that for $1 \leq T \leq n$) and dies out gradually. The occurrence of TC affects the time series for several lags depending upon the decay parameter δ with the size of outlier with magnitude ω_{TC}. This decay is sharper in ARMA(1,1) than MA(1). For Brevity purpose, we have excluded the detailed algebraic manipulations of these findings (for detail see Urooj; 2016 and Urooj& Asghar; 2017).

Following the impact of outliers on the autocorrelation function (ACF) of MA (1) and ARMA (1, 1) model; we have observed that in some cases the impact depends on the model specification as well along with the type of outlier while in other, it does not. The presence of AO makes the autocorrelation function downward biased. In the presence of large sized AO, the autocorrelation function is pushed toward zero. The impact of AO on any model does not involve model parameters and its specification. The sample ACF of MA (1) and ARMA (1, 1) process gets downward biased due to IO, however, the nature of bias depends upon the model parameters. The ACF approaches to zero for very large IO. The ACF with LS is upward biased. Large LS pushes the ACF toward unity. In the presence of TC, the ACF is upward biased and for large sized TC, it is dragged toward the decay parameter. However, for large sample size or small outlier size the bias in ACF reduces and the impact of all types of outliers fades away.

Studying the effects of different types of outliers on the estimates of coefficients, reveals that the observed series z_t having

MA(1) as data generating process will yield the MLE for θ_1 affected by the outliers in inverse proportion. The ARMA(1,1) yield estimation of additional parameter due to outliers. Exploring the impact of various outliers on the residuals of MA(1) and ARMA(1,1) model, the AO affects the residuals of MA(1) model up to few lags with the magnitude proportionate to the size of MA(1) coefficient. In ARMA(1,1) model, the AO affects the residuals up to few lags with the magnitude equals to the die down size of MA(1) coefficient with a constant value equals to the difference between AR(1) and MA(1) coefficients. The IO affects the residuals only at point 'T' with other residuals remain unaffected. This holds for both models. The LS affects the residuals at all points on and after 'T'. However, the impact on residuals with LS for MA(1) and for ARMA (1,1) model respectively are given as

$$e_t = \begin{cases} a_t & for\ t < T, \\ a_t + \omega_L\left(1 - \sum_{j=0}^{k} \theta_1\right) & for\ \mathrm{t} \geq T + k\ ; \\ & k = 1,2,\dots,(n-T) \end{cases} \tag{2}$$

and

$$e_t = \begin{cases} a_t & for\ t < T, \\ a_t + \omega_L[\ (1 + \sum_{j=1}^{k} {\theta_1}^j) - \phi_1(1 + \sum_{j=1}^{k} {\theta_1}^{j-1})] & for\ \mathrm{t} \geq T + k\ ; \\ & k = 1,2,\dots,(n-T) \end{cases} \tag{3}$$

The residuals for z_t with TC depends upon the decay parameter δ as $0 < \delta < 1$, i.e. closer the δ to 1, the slower is the decay and the TC behaves similar to LS. While if δ is closer to zero, the faster will be the decay and the TC gets closer to AO. For MA(1), the residuals form

$$e_t = \begin{cases} a_t & for\ t < T \\ a_t + \omega_{TC} \sum_{j=0}^{k} \delta^{k-j} {\theta_1}^j & for\ \mathrm{t} \geq T + k\ ; \\ & k = 0,1,2,\dots,(n-T) \end{cases} \tag{4}$$

And for ARMA (1,1), it is

$$e_t = \begin{cases} a_t & for\ t < T \\ a_t + \omega_{TC} & for\ t = T \\ a_t + \omega_{TC}(\sum_{j=0}^{k} \delta^{k-j} {\theta_1}^j - \phi_1 \sum_{j=0}^{k-1} \delta^{k-1-j} {\theta_1}^j) & for\ \mathrm{t} \geq T\ or\ t = T + j \\ & and\ j = 1,2,\dots n-T \end{cases} \tag{5}$$

3. Research Operationalization

As noted, different types of outliers affect the time series models in distinct form making it necessary to detect these outliers along with their type, magnitude and adjust them before any further analysis. For investigation of magnitude and dynamic effect of outliers on selected time series and the performance of test statistics for detection of outliers, we have used Monte Carlo simulation experiments on MA(1) and ARMA(1,1) process. The process is repeated for a total of 5000 iterations in CRAN-R for all combinations of the length of series(n): {50, 100, 150}, cut-off(C): {3,3.5,4},

Outlier size (ω): $\{3\sigma, 5\sigma\}$, parameters $for\ MA(1); \theta =$ {0.1, 0.2, 0.4, 0.6, 0.8, and 0.9} and for ARMA (1, 1): ($\phi = 0.7, \theta = 0.7$), ($\phi = 0.2, \theta = 0.8$), ($\phi = 0.4, \theta = 0.4$) and ($\phi = 0.8, \theta = 0.8$) are used. The analysis designed under hypothesis testing specifies null hypothesis (H_0) that no outlier is present and alternative hypothesis (H_1) as outlier/ structural break is present in the series. The analysis under H_1 is conducted for each type of break by calculating empirical power of test as relative frequency of correct detections. Then, the estimated 90th, 95th and 99th percentiles of the sampling distribution of test statistic η_i are calculated to provide insight of the patterns and behavior of the test statistics. We also follow the impact of outliers on estimation of parameters by following the sampling distribution of estimators of parameters. Lastly, correct index detection for the efficiency of the test statistics is noted. Under the null hypothesis (H_0), empirical level of significance is calculated as relative frequency of false outlier detection. Secondly, empirical critical values are calculated enabling us to determine if the three cut-offs(C) used are empirically valid.

4. Performance of Test Statistics in MA (1) and ARMA (1, 1) Model

The graphical view from Figure 1 indicates the impact of the four types of outliers on the simulated MA(1) and ARMA(1,1) processes affected by an outlier of magnitude $\omega = \{3, 5\}$at T= (n/2)+1 respectively. These outliers and their impacts, are located, identified, and estimated through the test statistics defined on likelihood ratio test criterion in section 3. Now we evaluate the performance of these statistics under different scenarios.

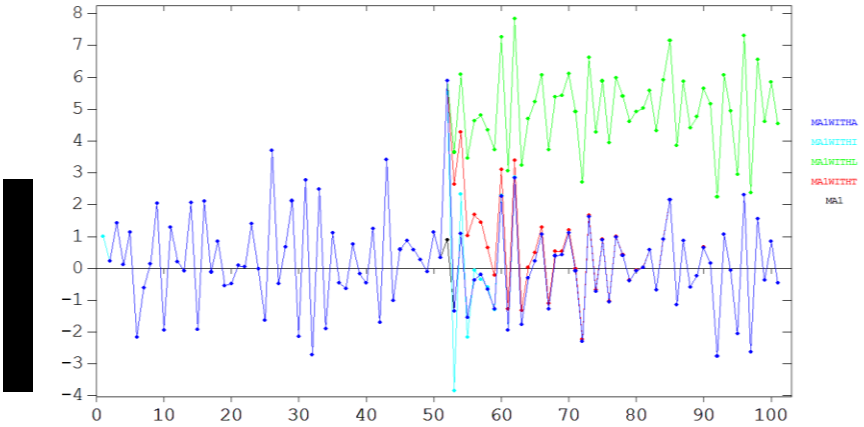

Figure1(a): MA (1) Model with and without Outliers

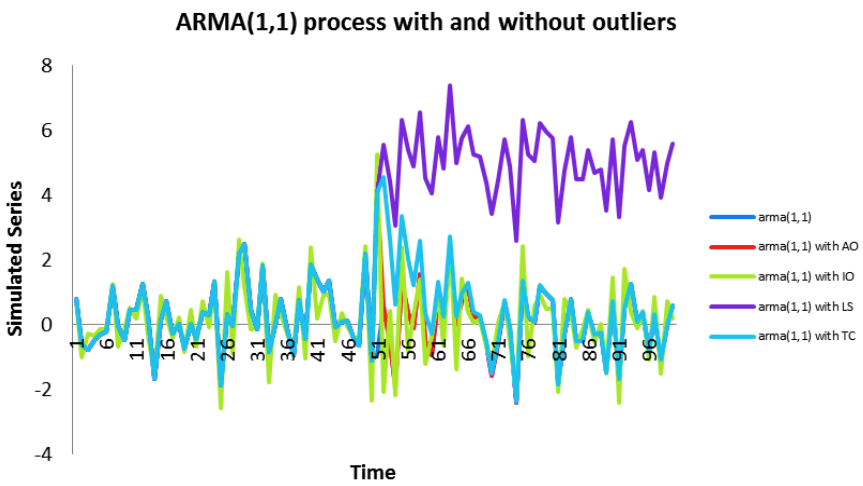

Figure1(b): ARMA (1) Model with and without Outliers

4.1. Sampling Distribution of Estimators in the Presence of Outlier

In order to study the impact of outliers on estimation, we observe the sampling distribution of estimators when MA (1) and ARMA (1, 1) models are affected by one outlier.

4.1.1. Sampling Distribution of $\hat{\theta}$; Case of MA (1) Model

The existence of various types of outliers in MA (1) process affects the sampling distribution variantly. AO causes downward bias in sampling distribution of $\hat{\theta}$ which increases with increase in θ values indicating if

MA(1) parameter has stronger impact to be carried to next lag; the AO has greater adverse effect. LS causes a constant bias irrespective of θ, n and C, such that $E(\hat{\theta}) \approx 0.92$ is noted for $\omega_{LS} = 5\sigma$ and in case of $\omega_{LS} = 3\sigma$ we get $E(\hat{\theta}) \approx 0.82$. TC causes a bias of varying nature. At small values of θ, the bias is small and positive while for large values of θ the bias is negative and large. In the presence of IO, the sampling distribution of $\hat{\theta}$ does not show any bias till θ=0.6 but for higher values of θ an upward bias is noted, even $E(\hat{\theta})$ >1 is observed at $\theta = 0.8, 0.9$ for all combinations of n, C and ω. The sampling distribution of $\hat{\theta}$ is non-normal in the presence of all outliers except for an IO at for θ>0.6 and n>100.

Negatively skewed sampling distribution of $\hat{\theta}$ is noted for LS, TC and IO but is positively skewed for AO in majority cases. Finally, mesokurtic sampling distribution of $\hat{\theta}$ is noted in the presence of IO, LS, and TC but leptokurtic in case of AO. The efficiency of various test statistics in terms of correct index detection of outliers, works satisfactory in case of all outliers except LS with $\omega_{LS} = 5\sigma$ and small cut-offs. Correct index detection at $\omega_{LS} = 5\sigma$ remains on average about 65% only. In case of all outliers of size $\omega = 3\sigma$ no good performance is noted (See Table 1(a, b, c, d)).

4.1.2. Sampling Distribution of θˆ; Case of ARMA(1,1) Model

The simulation exercise shows that AO, LS and TC causes huge downward bias, not much affected by the cut-offs(C), on sampling distribution of $\hat{\theta}$ in ARMA(1,1) model. The bias due to LS is not even affected by sample size; however, the bias due to AO reduces while that due to TC increases as the size of sample increases. In the presence of IO, the sampling distribution of $\hat{\theta}$ shows almost no bias for all combinations of n, C and ω. The RMSE and standard error (SE) of the sampling distribution of $\hat{\theta}$, in the presence of AO, IO, LS and TC, are not affected by C. Moreover, these remain unaffected by sample size under AO and LS, however, for IO these become decreasing function while in case of TC these become increasing function of sample size. The sampling distribution of $\hat{\theta}$ appears leptokurtic and non-normal in the presence of outlier of any kind except IO for $\phi = 0.8, \theta = 0.2$ at n= 150, C=3, $\omega = 5\sigma$ and $\phi = 0.4, \theta = 0.4$ at $n = 150$, $C = 3.5, \omega = 5\sigma$ only. The efficiency of test statistics in terms of correct index detection under the presence of AO, IO, LS and TC is very high

for $\omega = 5\sigma$ but poor for $\omega = 3\sigma$ which improves marginally as sample size increases (See Table 2).

4.1.3. Sampling Distribution of $\hat{\phi}$; Case of ARMA(1,1) model

The sampling distribution of $\hat{\phi}$ under ARIMA(1,1) process show downward bias in presence of AO, minor downward bias under IO, large upward bias of constant nature under LS and negligible upward bias under TC which remains unaffected by the cut-offs and outlier size. IO and LS remain unaffected by sample size while bias under AO and TC reduces as sample size increases. The sampling distribution of $\hat{\phi}$ is non-normal, leptokurtic and negatively skewed at all combination of n, C, and ω for all outliers with exceptions as with ($\phi = 0.4, \theta = 0.4$), n= 150, c= 3 for $\omega_{IO} = 5\sigma$ the sampling distribution does not yield significant JB results and for TC, the sampling distribution of $\hat{\phi}$ is mesokurtic. The SE and RMSE of sampling distribution are not affected by sample size, cut-off, and size of outliers under AO and LS while for IO and TC, these reduce as sample size increases and increase as cut-offs increases (See Table 3).

4.2. Empirical Level of Significance

Empirical level of significance is calculated as relative frequency of detection of any false outlier in an outlier free series.

4.2.1. Case of MA (1) model

The empirical significance level falls as θ increases but for high and moderately sensitive detections only. As the sample size increases, the empirical level of significance increases. As evident from Table 4, the increase in sample size causes more erroneous detections.

Table 4: Empirical Level of Significance for MA(1) Model

	C=3			C=3.5			C=4		
	n=50	n=100	n=150	n=50	n=100	n=150	n=50	n=100	n=150
MA(1)	T=26	T=51	T=76	T=26	T=51	T=76	T=26	T=51	T=76
θ =0.1	0.104	0.21	0.335	0.015	0.034	0.056	0.001	0.007	0.005
θ =0.2	0.128	0.225	0.3	0.021	0.038	0.066	0.003	0.004	0.007
θ =0.4	0.084	0.224	0.316	0.017	0.036	0.048	0.006	0.004	0.008
θ =0.6	0.105	0.207	0.304	0.015	0.036	0.055	0.003	0.002	0.002
θ =0.8	0.076	0.177	0.255	0.013	0.03	0.042	0.004	0.004	0.008
θ =0.9	0.08	0.191	0.271	0.007	0.021	0.043	0.002	0.004	0.008

When θ and C are small. The rate of change in level of significance due to a change in the sample size is very sharp. With the increase in sample size, at lower levels of C and θ, the empirical level of significance increases by more than 22% while it rises to 40% or 0.44 units for large values of C and θ indicating highly negative impact of sample size on test statistics' performance. In absolute terms as C is raised to a less sensitive point, the empirical level of significance falls remarkably and indicates better performance. In comparison with the nominal significance level (α), the empirical level of significance is higher and unsatisfactory at all θ and n for C=3. However, in less sensitive detections, the empirical level of significance reduces sharply and even falls below the nominal level of significance (α=0.05).

4.2.2. Case of ARMA(1,1) Model

Empirical level of significance shows interesting behaviour under ARIMA(1,1) model. It falls as n increases for parameter combinations ($\phi = 0.7, \theta = 0.7$) and ($\phi = 0.2, \theta = 0.8$) while it decreases with the increase in n for ($\phi = 0.4, \theta = 0.4$) and ($\phi = 0.8, \theta = 0.8$). This quantity also shows relation with C i.e. at less sensitive detection of outliers; it rises. We see, from Table 5, that the increase in sample size causes more variation when C is small. It attains less than 0.05 level at several combinations of C and sample size especially at C=3.5 indicating inefficient performance of test statistics for outlier detection at large cut-offs. The varying behaviour of empirical level of significance for different values of θ and ϕ show the dependency on model parameters.

Table 5: Empirical Level of Significance for ARMA(1,1) Model

	C=3			C=3.5			C=4		
	n=50	n=100	n=150	n=50	n=100	n=150	n=50	n=100	n=150
ARMA(1,1)	T=26	T=51	T=76	T=26	T=51	T=76	T=26	T=51	T=76
θ=0.4, φ=0.4	0.120	0.018	0.003	0.225	0.044	0.005	0.335	0.072	0.010
θ =0.7, φ =0.7	0.076	0.213	0.330	0.015	0.041	0.061	0.002	0.006	0.008
θ =0.2, φ =0.8	0.126	0.265	0.358	0.024	0.049	0.044	0.003	0.006	0.008
θ =0.8, φ =0.2	0.090	0.184	0.282	0.013	0.030	0.057	0.002	0.003	0.008

4.3. Empirical Critical Values

The empirical critical values, under different ARMA (p,q) processes, as suggested under simulation experiment, are higher than 3.

4.3.1. Case of MA (1) Model

The simulation exercise shows that the sampling distribution of false detection is more concentrated in MA(1) process than those under AR (1) model (Urooj and Asghar; 2017). Comparing the theoretical cutoffs with empirical critical values indicates that the cutoffs can be raised a little to get less false detections. However, the empirical critical values are not much influenced by the change in the magnitude of the parameter θ but show variation over theoretical cutoffs (See Table 6).

Table 6: On Average Empirical Critical Values for MA(1) Model

	90%	95%	99%
n=50	3.97	4.0706	4.2172
n=100	3.9178	4.0889	4.355
n=150	4.0833	4.245	4.505
average	3.9904	4.1348	4.3591

Note: Detailed tables of Empirical Critical values can be obtained on demand.

4.3.2 Case of ARMA (1, 1) model

Under ARMA (1, 1) model, the empirical critical values are influenced by the cut-offs, sample size and parameter values i.e θ and ϕ. For C=3; the empirical critical value on average lies around 3.8, for C=3.5 it is around 4.32 while at C=4, it is on average, more than 4.5(approx.) (See Table 7).

Table 7: Empirical Critical Values for ARMA(1,1) Model

Upper quantiles for false detection, c=3									
	n=50			n=100			n=150		
ARMA(1,1)	T=26			T=51			T=76		
	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
θ=0.4, φ=0.4	3.016	3.220	3.674	3.242	3.422	3.830	3.367	3.536	3.930
θ=0.7, φ=0.7	0.000	3.084	3.446	3.594	3.760	4.127	3.651	3.802	4.195
θ=0.2, φ=0.8	3.661	3.834	4.228	3.654	3.938	4.312	3.653	3.802	3.875
θ=0.8, φ=0.2	3.590	3.740	4.060	3.652	3.840	4.150	3.674	3.813	4.183

4.4. Power of Test Statistics η_j; $j = AO, IO, LS, TC$

Empirical power of the test statistics in the presence of a single outlier is studied as another yardstick of performance. It indicates the sensitivity of the statistical test in detecting changes (outliers) and is measured as relative frequency of rejecting the null of no outliers when in fact it is false.

4.4.1. Case of MA(1) Model

In case of AO, for $\omega_{AO} = 5\sigma$, the empirical rejection frequency is not satisfactory for all levels of θ, n and C. At C=3 and small θ values, the

power of η_{AO} $remains\ less\ than$ 50%. The power of η_{AO} increases remarkably as θ increases but shows very small improvement for n = 100 and falls at n=150. It shows high fluctuations in case of small sized outlier i.e. over 32% to 90%. The performance of η_{AO} for small AO remains less than 50% by and large. Misspecification frequencies indicating masking of outliers show that AO go largely unchecked as "no outliers" for small sized outlier. These cases are extremely high at C=3.5 and C=4 making the performance of η_{AO} questionable (See Table 8(a)). AO also confounds with IO and TC very frequently at all levels of ω, n, C and θ. However, the confusion with IO reduces many folds as θ increases.

For large IO ($\omega_{IO} = 5\sigma$), the empirical power of η_{IO} is relatively better than that of η_{AO}. Table 8(b) shows that the empirical power of η_{IO} varies over 53% to 97% at C=3 but reduces sharply for less sensitive cut-offs (C=4). In the case of Small IO ($\omega_{IO} = 3\sigma$), the empirical power of η_{IO} is very poor like that of η_{AO}. A rise in MA(1) parameter has a positive impact on empirical power of η_{IO}. For $\omega_{IO} = 5\sigma$, the increase in sample size has negligible impact on empirical power of η_{IO}, while it falls for small IO. Large IO is frequently detected as AO for small θ and large n. As n, C, and θ increases, the perplexity between AO and true IO fades away. Despite of the presence of IO, several iterations skip declaring "no outlier". This confusion reduces in case of large IO, small sample size, high cut-offs except at n=150 and large θ. Few erroneously detected TC are also observed (See Table 8(b)).

Performance of η_{LS} is very strong, attaining higher empirical power than η_{IO} and η_{AO}. With the increase in θ, the empirical power of η_{LS} increases, but for $\theta > 0.6$, it declines yet remains high. $\eta_{LS}$is not much affected by the sample size and cut-offs except for C=4 where it reduces to 65%(approx.). It performs poor in small sized outlier and gets even worse at n=50 and C=4. The LS is usually missed as "no outlier" i.e. all test statistics remains insignificant. However, for large sample size and sensitive cutoffs, only few cases of LS masked as TC, IO and AO are noted. The empirical power of η_{TC} is a function of sample size, θ, cutoffs and outlier size. For $\omega_{TC} = 5\sigma, \eta_{TC}$ performs well for large series, $C = 3$ and small θ, but the empirical power drops to very low with a rise in θ, increase in cutoffs and small outlier.

4.4.2. Case of ARMA(1,1) Model

Empirical power of η_{AO} for $\omega_{TAO} = 5\sigma$ in ARMA(1,1) shows satisfactory performance at C =3. It remains greater than 85% with small n and increases further for large n. As the cut-off increases, the empirical power of η_{AO} drops sharply. For small AO ($\omega_{AO} = 3\sigma$), the performance weakens showing high fluctuations and remains less than 78% by and large. AO gets mostly masked with IO, TC and is skipped as 'no outlier' for all ω_{AO} and large n (See Table 9 (a), Table 9(b)). The confusion with IO reduces as for large cut-off but show no impact of sample size, with 'no outliers' at C=3 and the confusion with TC does not show clear relation with n and C. The empirical power of η_{IO} is satisfactory for $\omega_{IO} = 5\sigma$ at C=3 only. It falls as low as 7% for small ($\omega_{IO} = 3\sigma$) which is undesirable. Negative impact of increase in cut-off(C) and little impact of increase in sample size on empirical power of η_{IO} is noted IO is largely confused with TC or is missed out as "no outliers". This confusion increases as C and n increases and for small sized outlier. Empirical power of η_{LS} is low at all sample sizes and is not much affected by the sample size and cut-off. It is mostly confused with "no outliers" along with only few instances of misidentification as TC, IO and AO. The empirical power of η_{TC} is a function of sample size, performing well for large series. TC is highly masked as IO even for large outlier. It is also perplexed with AO. For small sized TC($\omega_{TC} = 3\sigma$), the performance of η_{TC} is not impressive at all combinations of n and C.

4.5. Behavior of δ in Transient Change

The performance of η_{TC} for various choices of δ has also been studied. In MA(1) process, for extreme values of δ, η_{TC} performs very poor. At $\delta = 0$, η_{AO} and η_{TC} yield exactly same values in almost 30% of iterations. Secondly, instead of detection of TC, IO has been detected frequently and there are some cases of "no outlier" identified at all combinations of parameters. Unlike the confusion with AO, the erroneous detections as IO and "no outliers", show negative association with sample size and θ. As δ value is raised to some non-zero number, the confusion between η_{AO} and η_{TC} vanishes off, but no significant decrease in the number of erroneously detected IO and "no outlier" cases. The percentage of correct detections of TC increases gradually with an increase in δ. At $\delta = 0.6$, for the combinations {n=50, $\theta = 0.1$}, and {n=150, $\theta = 0.1, 0.2$}; η_{TC} attains an empirical rejection frequency of about 80% or more. Beyond these values of θ, the

empirical power falls very sharply even below 50%. Improved performance of η_{TC} is observed for sensitive detections at $\delta = 0.8$ with small values of θ and large n. Similarly, high empirical power is also noted at $\delta = 0.9$ for all values of θ and n. However, at δ=1, the power of η_{TC} drops to zero with $\eta_{TC} = \eta_{LS}$ in several iterations. Here "no outlier" cases are also too many. We conclude that the test statistics of η_{TC} for detecting TC performs adequat only for the choices of $\delta = 0.8$ and $\delta = 0.9$. The confusion between η_{TC} and η_{IO} noted throughout the simulations must also be considered whenever an outlier detection procedure is applied in practice. (See Table 10(a, b)).

5. Empirical Analysis

Structural changes in various frontiers are contributed not only due to variations in several factors but also due to many unanticipated events like floods, earthquakes, epidemics, large scale energy crises etc. These structural changes appearing in form of outliers and structural breaks are in keen interest of policy makers and researchers. To gain insight, the exploration of outliers for Pakistan provides an ideal case study. The empirical study conducted on some monthly time series for Pakistan is applying three-stage outlier detection procedure suggested by Chen and Liu (1993). The iterative testing procedure provides consistent estimates of the model covering entire sample as well as consistent estimates of the true number of breaks (for details see Urooj; 2016 and Chen & Liu; 1993). This procedure is based upon the test statistics whose performance is observed in earlier sections. The analysis with structural breaks for Pakistan, proves useful and provides evidence of the presence of variations in several social, environmental, geographical and economic aspects which needs to be taken into consideration when modeling the macroeconomic, socio and ecological growth nexus for policy makers.

We examine 5 monthly measured time series of Pakistan (See Figure 2). Two of these series span over February1995 to February 2015, the gold prices span over December 2000 to February 2015, while the net effective exchange rate extends over January1980 to February2015 and KSE-100 closing prices range over June1994 to September2020 due to availability of data (See Table11).The data are taken from IFS, World Bank, SBP reports and metrological department.

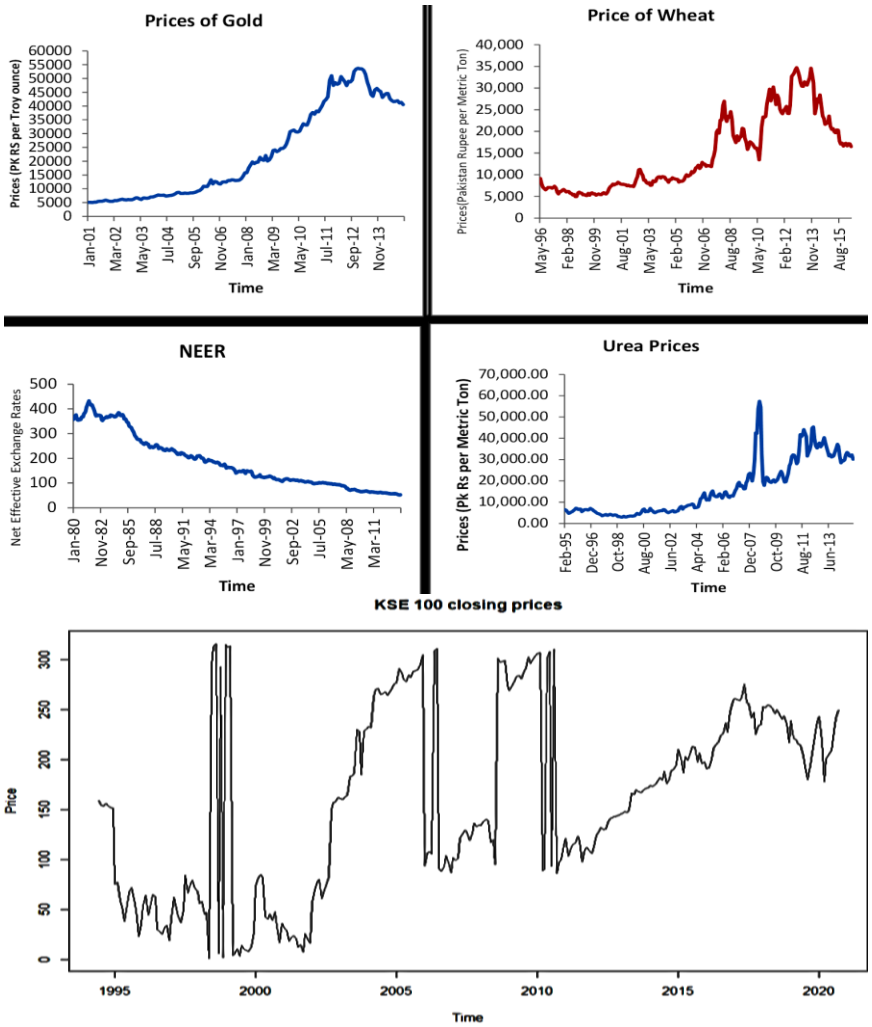

Figure 2: Monthly Time Series Data with Outliers

Looking at Table 11 for the descriptive statistics for the variables under study indicates large variations over full range of data. The skewness statistic indicates positive asymmetric behavior of all variables except in gold prices meaning that during sample period, there were more decreases in gold prices than large increases. In addition, negative excess kurtosis results in significant JB indicating a non-Gaussian distribution, while Ljung-Box Q statistics test for autocorrelation up to 24 lags indicates existence of autocorrelation in all series. These series also indicate

existence of annual unit root with no requirement for seasonal differencing. Hence, an ARIMA process to capture the dynamic structure and to generate white-noise residuals is suggested.

5.1. Outlier Detection and Intervention Model

The results in form of parameter estimates, their standard errors, residual's standard errors, skewness and kurtosis of residuals for with and without outlier detection are listed in Table 12. All the series show significant evidence of excess kurtosis and skewness. The JB test for all series indicates that the initial model estimation (without outliers) generates non-Gaussian residuals. However, it falls remarkably in 'with-outlier' analysis for all series even supports the possibility of Gaussian residuals in case of NEER. We generalize that outliers if neglected may lead to excess kurtosis, skewness of residuals and significant JB test making standard statistical theory based on Gaussian distribution redundant. Comparing the results of initial identified model with those obtained incorporating the outliers via Three-stage Chen and Liu (1993) procedure under Table 12 shows that the error variance of the originally identified model is greater than that under the intervention analysis. Not only the values of estimated parameters change but also the standard error of all estimated parameters reduces. The results under the Chen and Liu procedure for joint estimation of outliers give significant evidence (at C=3.5) of outliers and structural breaks in all series. These identified outliers explain substantial proportion of volatility in majority of the series. The proportion of variation explained by outliers, as suggested by Balke and Fomby (1994) is calculated as:

$$Proportion\ of\ variation\ explained\ by\ outliers = 1 - \left[\frac{var(\hat{y})}{var(y)}\right] \quad (6)$$

where $\widehat{y_{out}}$ is the fitted value obtained from the model with outliers and y is the observed series.

The last column of Table 12 gives the proportion of variation explained by outliers. It is evident that the outliers for a series may have explained as low as 6.02% of total variation and for some other series may have explained as high as 39.77% of the total variation. In case of NEER, the proportion of variation explained due to outliers is about 6.02% and for gold prices it is 39.77%, highlighting the importance of

outliers/ breaks in explaining the dynamics of the time series under study. Thus, the identification and estimation of possible outliers along with other features of time series are very important and necessary.

Among the outliers identified, AO are most common, several LS are detected while few TC are identified. In general, clustering of outliers within series are noted at several instances where different series have outliers at or near the same date. The types of outliers in these cases may or may not match. Table 13 presents the outliers found in each series, their type and size as well as the date at which they have occurred in chronological order, as it is more viable to observe the patterns of outliers across time and series. There appear to be a clustering of outliers within series. It is well documented in literature that majority of outliers are associated with business cycle, particularly recession. However, since all business cycles are not same, so, is the behavior of outliers during these cycles requiring in-depth analysis. These joint occurrences also point toward the possibility of some political and economic events occurring in the country. The clustering of outliers across time is evident in Urea prices series at 2004 (June, July) then in 2004 (November, December). Within the clusters of outliers, the types of outliers identified may vary as in Urea prices the two outliers at June and July 2004 are AO and IO while the successive outliers at November and December 2004 are identified as LS and TC. The theory of outlier detection postulates LS to be a permanent break while TC is the break that decays off. Their successive existence requires exploration as this may be an issue of biased initial model identification or an incidence of misidentification as the graphical view of the series indicates possible LS. It may be a shortfall in the Chen and Liu procedure as mentioned by Sanchez and Pena (2003) but needs further exploration. Hence, along with the statistical exploration of the issue, evidence between the occurrence of these outliers and their historical perspective may prove helpful. The outlier detection procedure when applied to the monthly data of wheat results in two significant outliers i.e. AO at June 2010 and LS at July 2010. The intervention model for price of wheat for initially detected model as SARIMA $(0,1,1)(0,0,0)_{12}$ is written as

$$\nabla \ln(p_{wheat}) = a_t - 0.2826a_{t-1} - 0.1426I_t^{2010M_6} + 0.2108\left(\frac{1}{1-B}\right)I_t^{2010M_7} \quad (7)$$

Similarly, an intervention model for gold prices with three AOs can be written as

$$\nabla \ln(p_{gold}) = a_t - 0.4152a_{t-1} - 0.1077I_t^{2006M_5} + 0.0680I_t^{2008M_7} - 0.0949I_t^{2008M_{10}} \quad (8)$$

The intervention models for other series are listed in Table 13.

Table 13: Details of Detected Outliers from Monthly Time Series of Pakistan

Variable	Model	Index	Time	Estimate	Std. error	t-value	Type
ln(gold pk)	(0,1,1)(0,0,0)+C	66	05,2006	0.10774	0.0181	5.967	AO
		92	07,2008	0.06801	0.0183	3.717	AO
		95	10,2008	0.09492	0.0183	5.200	AO
ln(urea)	(0,1,1)(0,0,0)	35	12,1997	0.4231	0.0999	4.237	AO
		113	06,2004	1.4855	0.1507	9.859	AO
		114	07,2004	-0.5468	0.0946	-5.778	IO
		116	09,2004	1.1810	0.2258	5.230	TC
		118	11,2004	-1.1180	0.248	-4.508	LS
		119	12,2004	-1.6307	0.2258	-7.222	TC
		122	03,2005	3.3939	0.1963	17.291	LS
ln(Wheat)	(0,1,1)(0,0,0)	132	01,2006	-0.8673	0.1006	-8.619	AO
		141	10,2006	-0.4467	0.0997	-4.482	AO
		185	06,2010	-0.1426	0.0369	-3.868	AO
		210	07,2010	0.2108	0.0577	3.652	LS
ln(NEER)	(0,1,1)(0,0,0)+C	191	11,1995	-0.05334	0.013	-4.017	LS
		203	11,1996	-0.04812	0.012	-4.13	TC
		224	08,1998	-0.05358	0.015	-3.681	IO
		233	05,1999	0.05998	0.008	7.588	AO
ln(KSE100)	(0,1,1)(0,0,0)	8	01,1995	-0.7757	0.319	-2.4284	LS
		18	11,1995	-0.2656	0.241	-1.1016	IO
		31	12,1996	-0.7157	0.314	-2.2801	AO
		48	05,1998	-4.7242	0.315	-15.0051	AO
		52	09,1998	-0.5599	0.289	-1.9368	IO
		54	11,1998	-4.3819	0.320	-13.6753	AO
		58	03,1999	-4.0913	0.335	-12.2062	LS
		60	05,1999	0.1555	0.266	0.5851	IO
		62	07,1999	0.4774	0.387	1.2342	TC
		78	11,2000	-0.2225	0.241	-0.9251	IO
		88	09,2001	-0.9716	0.314	-3.0907	AO
		92	01,2002	1.2655	0.319	3.9672	LS
		140	01,2006	-1.0497	0.320	-3.2779	LS
		171	08,2008	1.0162	0.320	3.1805	LS
		190	03,2010	-0.6058	0.243	-2.4937	IO
		195	08,2010	0.9854	0.315	3.1311	AO

6. Conclusions

The extensive simulation experiment identifies that the sampling distributions of estimators for the parameter of contaminated series are biased, skewed and non-normal. The outliers need to be large for the method to have decent power. For small sized outliers, η_{LS} give average performance while other test statistics show poor performance. For sensitive detections (C=3), the empirical level of significance is

higher than the nominal level of significance; selection of slightly higher cutoffs(C) may help in reducing the chances of false detections. However, large cutoffs as identified under null hypothesis are not much supported in terms of power of the test statistics. Misspecifications among AO, IO and TC are also observed. The skipping in form of "no outlier" indicates the weakness of test statistics and appears frequently large cutoffs and for small outliers. Hence, outlier size needs to be large to have good performance of statistics. The decaying parameter should be used as high as 0.85 or 0.9 or δ should be estimated via some nonlinear estimation technique for satisfactory performance of test statistics of TC. This indicates that there is need to revisit the test statistics for TC and IO.

The empirical analysis has shown that neglecting the presence of outliers affects the identification, estimation and results in poor statistical analysis. The detection and removal of outliers and structural breaks reduces the residual's excess kurtosis, skewness and JB test remarkably. The analysis has identified several statistically significant shocks in all series under study. The possibility of incidence of misidentification, masking and swamping effects in identified outliers needs further exploration. It seems important to use the critical information being translated in these indicators in form of outliers and structural breaks. Connecting the indicated discordant observations with historical evidences helps in better understanding of past policies and designing effective policies in future.

Conflict of Interest	None
Supplementary Martial	No supplementary material is associated with the article
Funding	This research received no external funding
Acknowledgment	I, Amena Urooj, hereby declare that this paper is a part of my thesis.
ORCID of Corresponding Author	https://orcid.org/0000-0003-4626-5525

References

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control, revised ed*. Holden-Day. Retrieved from https://www.amazon.com/Time-Analysis-Forecasting-Control-Hardcover/dp/B011SJ22JI.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*(2), 193-204. https://doi.org/10.1080/00401706.1988.10488367.

Chen, C., & Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*(421), 284-297. https://doi.org/10.1080/01621459.1993.10594321.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(3), 350-363. https://doi.org/10.1111/j.2517-6161.1972.tb00912.x.

Kaiser, R., & Maravall, A. (2001). Seasonal Outliers in Time Series, Estadıstica. *Journal of the Inter-American Statistical Institute*, *53*, 101-142.

Pena, D. (1990). Influential observations in time series. *Journal of Business & Economic Statistics*, *8*(2), 235-241. https://doi.org/10.1080/07350015.1990.10509795.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, *81*(393), 132-141. https://doi.org/10.1080/01621459.1986.10478250.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, *7*(1), 1-20. https://doi.org/10.1002/for.3980070102.

Urooj, A. (2016). *Performance of Time Series Models under Structural Discontinuities and Discordant Observations*. [Unpublished Ph.D. thesis]. Quaid-i-Azam University.

Urooj, A., & Asghar, A. (2017). Analysis of the performance of test statistics for detection of outliers (additive, innovative, transient, and level shift) in AR (1) processes, *Communications in Statistics-Simulation and Computation*, 46(2), 948-979. https://doi.org/10.1080/03610918.2014.985383.

Citation: Urooj, A., & Asghar. Z. (2020). Evaluation of test statistics for detection of outliers and shifts, *Journal of Quantitative Methods, 4(*2), 54-75. https://doi.org/10.29145/2020/jqm/040203.