CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Serkan ÖCAL

DIAGNOSTIC METHODS FOR OUTLIER DETECTION IN TIME SERIES

DEPARTMENT OF STATISTICS

ADANA, 2020

CUKUROVA UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DIAGNOSTIC METHODS FOR OUTLIER DETECTION IN TIME SERIES

Serkan ÖCAL

MSc THESIS

DEPARTMENT OF STATISTICS

We certify that the thesis titled above was reviewed and approved for the award of degree of the Master of Science by the board of jury on 16/03/2020.

.....	.....	.....
Prof. Dr. Mahmude Revan ÖZKALE	Prof. Dr. Selahattin KAÇIRANLAR	Assoc.Prof.Dr. Özge KURAN
SUPERVISOR	MEMBER	MEMBER

This MSc thesis is written at the Department of Institute of Natural And Applied Sciences of Cukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Director
Institute of Natural and Applied Sciences

Note: The usage of the presented specific declaration's, tables, figures, and photographs either in this thesis or in any other reference without citation is subject to "The law of Arts and Intellectual Products" number of 5846 of Turkish Republic.

ABSTRACT

MSc THESIS

DIAGNOSTIC METHODS FOR OUTLIER DETECTION IN TIME SERIES

Serkan ÖCAL

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Mahmude Revan ÖZKALE
Year: 2020, Pages: 93
Jury : Prof. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Selahattin KAÇIRANLAR
: Assoc. Prof. Dr. Özge KURAN

It is likely to encounter anomaly observation, which stays significantly away from the rest of the data, in observed time series. This kind of observation is called as outlier and it usually occurs as a result of unpredictable events.

It is known that the existence of outlier in time series effects the parameter estimation and inflates the variance according to the outlier's type, thus causes misidentification of the model and poor forecasting. Because of these reasons, it is important to detect, if any, the outliers in time series.

In this study, detection of outlier in ARIMA time series models is considered. Iterative identification procedure is revised to correctly determine the time and the type of outlier. A simulation study is carried to investigate the power of the procedure by using RStudio. It is offered some suggestions for the pros and cons of the procedure.

Key words: Time series, ARIMA models, Additive outlier, Innovative outlier, Level shift

ÖZ

YÜKSEK LİSANS TEZİ

ZAMAN SERİLERİNDE SAPAN DEĞER TESPİTİ İÇİN TANILAMA METOTLARI

Serkan ÖCAL

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Mahmude Revan ÖZKALE
Yıl: 2020, Sayfa: 93
Jüri : Prof. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Selahattin KAÇIRANLAR
: Dr. Öğretim Üyesi Özge KURAN

Gözlemlenen zaman serilerinde tipik olmayan, verinin genelinden uzak duran gözlemlerle karşılaşılması muhtemeldir. Bu tür gözlemler sapan değer olarak adlandırılır ve genellikle beklenmedik olayların neticesinde ortaya çıkarlar.

Zaman serilerinde sapan değer varlığının parametre tahminini etkilediği ve serinin varyansını artırdığı, böylelikle modelin yanlış belirlenmesine ve güvenilir olmayan öngörüye neden olduğu bilinmektedir. Bu nedenle zaman serisinde sapan değer tespiti önemli bir yer tutmaktadır.

Bu çalışmada, ARIMA zaman serisi modellerinde görülen sapan değerler incelenmiştir. Sapan değerin seri içerisinde meydana geldiği noktayı ve türünü doğru bir şekilde tespit etmek için iteratif tanımlama yöntemleri gözden geçirilerek derlendi. RStudio programı kullanılarak yöntemin gücünü incelemek için simülasyondan üretilen zaman serisi üzerinde çalışıldı. Sapan değer tanılama prosedürünün artı ve eksileri için bazı önerilerde bulunuldu.

Anahtar Kelimeler: Zaman serileri, ARIMA modelleri, Anlık Sapan Değer, Yenilik Sapan Değer, Düzey Kayması

GENİŞLETİLMİŞ ÖZET

Düzenli zaman aralıkları ile elde edilen veriye zaman serisi denilmektedir. Ekonomide günlük dolar kuru, tarımda aylık ortalama üretim, biyolojide denekten alınan anlık kalp atış oranı, işletmede haftalık stok miktarı ve meteorolojide belli bir bölgeye düşen yıllık yağış oranı gibi birçok alanda zaman serisi verisi ile karşılaşılmaktadır. Zaman serisi elde edinildiği zaman aralığına göre saatlik, günlük, haftalık, aylık, yıllık vb. zaman serisi olarak adlandırılır.

Gözlemlenen zaman serisinde diğer gözlemlerden önemli ölçüde ayrı duran ve tahmin modeline uyum sağlamayan aykırı gözlemler ile karşılaşılması muhtemeldir. Bu aykırılığın sebebi kriz, savaş, grev, ölçüm cihazındaki hatalar vb. gibi önceden tahmin edilemeyen beklenmedik olaylar olabilir.

Zaman serilerinde sapan değer olması durumunda parametre tahmininin yanlı olması ve serinin varyansının artması sorunlarıyla karşılaşılmaktadır. Bu durum modelin yanlış belirlenmesi ve güvenilir olmayan öngörüyle sonuçlanmaktadır. Bu nedenle zaman serisinde sapan değer tespiti ve etkilerinin giderilmesi doğru model ve parametre tahmini elde etme ve böylelikle güvenilir öngörü yapabilme açısından önemli bir yer tutmaktadır. Yaklaşık son yarım asırdır bu konuyla ilgili çalışmalar yapılmaktadır.

Zaman serilerinde sapan değerler yapısı bakımından dört farklı türde görülmektedir. Bunlar; Anlık Sapan Değer (*Additive Outlier*-AO), Yenilik Sapan Değer (*Innovational Outlier*-IO), Düzey Kayması (*Level Shift*-LS) ve Geçici Değişim (*Temporary Change*-TC)'dir.

Fox (1972) AO türü sapan değeri tek bir gözlemi etkileyen bir gözlemdeki veya kayıtta yapılan büyük bir hata olarak ve IO türü sapan değeri meydana geldiği andan itibaren diğer gözlemleri de etkileyen gözlem olarak tanımlamaktadır. AO türü sapan değerin diğer sapan değer türlerine kıyasla modelin parametre ve varyans tahminlerini daha fazla etkilediği bilinmektedir. Tsay (1988) LS türü sapan değeri seride meydana geldiği noktadan itibaren düzeyi değiştirdiği ve bu

değişimin seri boyunca kalıcı olduğu şeklinde ve TC türü sapan değeri ise serinin ortalamasının kısa süreliğine değiştiği ve bu değişim etkisinin üssel olarak azalarak kaybolduğu şeklinde tanımlamıştır.

Sapan değerin, eğer varsa, zaman serisi içerisinde meydana geldiği noktayı ve türünü belirlemeye yönelik olarak Fox (1972) olabilirlik oran testini önerdi. Chang (1982) ve Chang ve Tiao (1983) Fox'un testini ARIMA modellerinde sapan değer tespiti için iteratif bir prosedüre genişletti. Daha sonra bu iteratif prosedür Chang ve ark. (1988), Chen ve Liu (1993) ve Ljung (1993) tarafından çeşitli düzenlemeler yapılarak geliştirildi. Pena ve Sanchez (2003) ise önceki yöntemlerin gücünü kaybetmeden başlangıç parametre tahminini ve sapan değer türü LS iken IO olarak tanılanmasını önleyecek şekilde iteratif tanılama yönteminde düzenleme yaparak geliştirdiler.

Zaman serilerinde bir gözlem sapan değer olarak belirlenmediği halde parametre tahminleri üzerinde etkili olabilir. Bu etkili gözlemleri tespit etmek için önerilen yöntemler genel olarak regresyonda etkili gözlemi tespit etmek için kullanılan yöntemlerin zaman serisi modelinde uyarlanması şeklindedir. Bruce ve Martin (1989) bir gözlemin etkinliğini ölçmek için gözlemin zaman serisinden silindikten sonra parametre tahminindeki değişimi ölçmeye dayalı tanılama ölçütü önerdi. Abraham ve Chuang (1989) regresyonda kullanılan Draper ve John (1981)'in önerdiği gözlemin silinmesine dayalı etkili gözlemleri belirleme istatistiği önerdi. Pena (1990) ARIMA modellerindeki katsayılar üzerinde etkili olan gözlemleri çalıştı ve bu etkili gözlemleri belirlemek için Mahalanobis uzaklığına dayalı etki istatistiği tanımladı. Hui ve Lee (1992) regresyon analizinde yaygın olarak kullanılan birini-dışarda-bırak (leave-one-out) etkili gözlem tanılama ölçütünü ARIMA modeline uyarladılar. Lin ve Lee (2001) ise Lawrance tarafından önerilen ve regresyonda kullanılan koşullu Cook uzaklığını zaman serilerine uyarladılar.

Bu tez çalışması beş bölümden oluşmaktadır. İlk bölümde, zaman serilerinde sapan değer tespitinin önemi hakkında bilgiler verildi. Bu konuda yapılan çalışmalar gözden geçirildi.

İkinci bölümde, zaman serisinin bileşenlerinden bahsedildi ve zaman serisi analizinde önemli yer tutan durağanlık kavramı üzerinde duruldu. Zaman serisinin durağan olup olmadığını belirlemek için yaygın olarak kullanılan birim kök testlerine yer verildi. Ardından literatürde Box ve Jenkins modelleri olarak da bilinen doğrusal zaman serisi modelleri olan AR, MA, ARMA ve ARIMA modelleri tanıtıldı ve bu dört modelin yapısal özellikleri teorik yapısı bakımından incelendi.

Tezin üçüncü bölümünde, genel olarak yukarda belirtilen çalışmalardan yararlanılarak zaman serilerinde sapan değerler incelendi. AO, IO, LS ve TC türlerine ait modeller tanıtılarak bu dört sapan değer türüne ait bilgiler verildi. Daha sonra etkili gözlemleri belirlemeye yönelik önerilen istatistikler tanıtıldı. Sapan değer belirleme yöntemi için yaygın olarak kullanılan iteratif tanılama prosedürü önceki çalışmaların bir birleşimi olarak derlendi ve sunuldu.

Dördüncü bölüm olan uygulama bölümünde, RStudio programı kullanılarak iteratif tanılama prosedürünün farklı türdeki sapan değerleri belirlemedeki etkinliğini incelemek için simülasyon ile üretilen seriye farklı türlerde sapan değerler eklendi. Seride sapan değerler varken seriye uydurulan model ile seriden sapan değerlerin etkisi giderildikten sonra uydurulan model arasında ciddi fark olduğu gözlendi. Buradaki farktan kasıt, sapan değerlerin etkisi seriden giderildikten sonra elde edilen seriye uydurulan modelin farklı olması, modelin parametrelerine ait standart hataların daha küçük olması, tahmin varyansındaki azalma ve uygun model belirlemede kullanılan AIC/BIC kriterlerindeki azalmadır. Bu bölümde ayrıca sapan değer tanılama prosedürünün seride ardışık farklı iki sapan değer olması durumunda hatalı sapan değer belirleme durumuna dikkat çekmek için simülasyon serisine eklenen sapan değerlerin seri içerisindeki yeri değiştirilerek prosedür uygulandı ve bulgular sunuldu. Yine aynı

simülasyon verisine farklı büyüklükte sapan değer eklenmesi sonucu, zaman serisi grafiğinde açıkça sapan değer olarak yorumlanabilen noktalar varken tanılama prosedürünün sapan değer belirleyemediği durumu ile karşılaşıldı.

Beşinci ve son bölüm olan sonuç bölümünde ise iteratif tanılama prosedürünün etkinliği hakkında çıkarımlar yapıldı. Uygulamada karşılaşılan ve sapan değer tanılama prosedüründe eksiklik olarak gördüğümüz durumlara yönelik prosedürde yapılabilecek iyileştirmeler için öneriler sunuldu.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Mahmude Revan ÖZKALE ATICIOĞLU for her patient guidance and advices. I appreciate the expertise and time that she has contributed throughout the course of this thesis.

I would like to thank my jury members, Prof. Dr. Selahattin KAÇIRANLAR and Assoc. Prof. Dr. Özge KURAN for their valuable suggestions and comments.

I would like to thank all staff of Department of Statistics at Cukurova University and my classmates.

My deepest gratitude goes to my family; my parents Sevda and Mehmet for their continuous love and confidence in me, my sisters Aslı, Aysel, Bahar and brother Ali for their encouragements and moral support throughout my life.

I am grateful to my friends M. Ersin GEZER and Muharrem KÜTÜK, who helped me tremendously and always motivated me.

Last but not least, I would also like to acknowledge my colleagues, who have directly or indirectly, knowingly or unknowingly contributed towards this thesis.

CONTENTS **PAGE**

ABSTRACT I
ÖZ II
GENİŞLETİLMİŞ ÖZET III
ACKNOWLEDGEMENTS VII
CONTENTS VIII
LIST OF FIGURES X
LIST OF TABLES XII
LIST OF ABBREVIATIONS XIV
1. INTRODUCTION 1
1.1. Literature Survey on Outliers in Time Series 2
2. TIME SERIES 5
2.1. Components of a Time Series 5
2.2. Some Simple Operators 7
2.3. Autocovariance and Autocorrelation Functions 8
2.4. Concept of Stationary 8
2.4.1. Significance Test for Autocorrelation Coefficient 11
2.4.2. Unit Root Tests 14
2.4.2.1. The Dickey-Fuller Test 14
2.4.2.2. Augmented Dickey-Fuller Test 16
2.5. Time Series Models 17
2.5.1. Stationary Time Series Models 18
2.5.1.1. Autoregressive Process 18
2.5.1.1.(1). A pth-order autoregressive process 19
2.5.1.1.(2). A first-order autoregressive process 22
2.5.1.1.(3). A second-order autoregressive process 26
2.5.1.2. Moving Average Process 29
2.5.1.2.(1). A qth-order moving average process 29

2.5.1.2.(2). A first-order moving average process 31
2.5.1.2.(3). A second-order moving average process 33
2.5.1.3. Invertibility 36
2.5.1.4. Mixed Autoregressive Moving-Average Process 38
2.5.1.4.(1). An ARMA(1,1) process 41
2.5.2. Non-Stationary Time Series Models 44
2.5.2.1. Autoregressive Integrated Moving Average Process 44
2.6. Model Selection 45
3. OUTLIERS in TIME SERIES 47
3.1. Additive Outlier (AO) 47
3.2. Innovational Outlier (IO) 49
3.3. Level Chance (LC) 50
3.4. Transient Chance (TC) 51
3.5. The General Outlier Model 53
3.6. Estimation of Outlier Effects 54
3.7. Testing for an Outlier 56
3.7.1. Confusion between IO and LC 57
3.8. Iterative Procedure for Outlier Detection 59
3.9. Influential Observations 61
4. APPLICATION 67
4.1. Outlier Detection with a Simulated Series 67
5. CONCLUSIONS 77
REFERENCES 79
CIRRICULUM VITAE 83
APPENDICES 84

LIST OF FIGURES **PAGE**

Figure 2.1. US Treasury bill contracts data plot 6
Figure 2.2. US Birth data plot 7
Figure 2.3. Pork retailer prices in Germany 7
Figure 2.4. Plots of simulated AR(1) series with different phi values 24
Figure 2.5. Plot of simulated random walk series 25
Figure 2.6. Plots of simulated AR(2) series 29
Figure 2.7. Plots of simulated MA(1) series with different theta values 33
Figure 2.8. Plots of simulated MA (2) series 35
Figure 2.9. Plot of simulated ARMA (1,1) series 43
Figure 3.1. Plot of simulated AR(1) with an AO at t=75 48
Figure 3.2. Plot of simulated AR(1) with an IO at t=75 50
Figure 3.3. Plot of simulated AR(1) with an LC at t=75 51
Figure 3.4. Plot of simulated AR(1) with an TC at t=70 52
Figure 4.1. Plot of simulated ARIMA (1,0,1) with phi=.5 and theta=-.4 67
Figure 4.2. Contaminated time series with AO at t=15,33 and a LC at t=110 68
Figure 4.3. Contaminated time series with adjusted series 72
Figure 4.4. Contaminated time series with LC shifted from t=110 to t=100 73
Figure 4.5. Contaminated time series with only TC of size 4 at t=105 till t=113 ... 75
Figure 4.6. Contaminated time series with only TC of size 5 at t=105 till t=113 ... 75

LIST OF TABLES **PAGE**

Table 3.1. Recommended critical values for the outlier type of IO and LS by Sanchez and Pena (2003) 59
Table 4.1. Estimated parameters for the simulated time series 68
Table 4.2. ADF test result for contaminated time series 69
Table 4.3. Fitted ARIMA to the contaminated time series 69
Table 4.4. Types and locations of detected possible outliers 70
Table 4.5. Types and locations of detected outliers 71
Table 4.6. Fitted ARIMA to the adjusted time series 72
Table 4.7. Detected outliers for contaminated series that LC shifted t=100 74

LIST OF ABBREVIATIONS

ACF	: Autocorrelation Function
ADF	: Augmented Dickey Fuller
AIC	: Akaike Information Criterion
AO	: Additive Outlier
AR	: Autoregressive
ARIMA	: Autoregressive Integrated Moving Average
ARMA	: Autoregressive Moving Average
BIC	: Bayesian Information Criterion
C	: Critical Value
CL	: Confidence Level
DF	: Dickey Fuller
IO	: Innovational Outlier
LS	: Level Shift
MA	: Moving Average
MAD	: Median Absolute Deviation
PACF	: Partial Autocorrelation Function
TC	: Temporary Change

1. INTRODUCTION

It is common to encounter anomaly observation, which stays considerably away from the remainder of observations, in the data set. Observation of this kind is usually called as outlier.

In literature, many definitions have been done to define outlier. Edgeworth (1887) writes: Abnormal points can be defined as those which present the appearance of differing in regard to their law of frequency from other points with which they are combined. 82 years later Grubbs (1969) says that: An outlier may be defined as one that seems to diverge significantly from the other observations (Beckman and Cook, 1983).

Barnett and Lewis (1994) had defined outlier as: an observation that appears to be inconsistent with the rest of that set of data, follow deviant view and can identify as being discordant with various statistical tests (Hodge and Austin, 2004).

There exists also some other naming for outlier such as aberrant, abnormal, discordant observations, anomalies, novelties, deviants, contaminants, and extreme point in the statistics literature.

There might be different causes to make outlier occurred in data, main important two of them are natural randomness and fault of measuring. Former one might be a natural disaster or a crisis in economy, and later one might be a fault of the person or measuring set, for instance, a measurement error of tool, or a mistake due to the lack of personal attention, or a systematic fault in program, etc.

The data have been used for predicting and forecasting, since outliers prevent them from obtaining accurate results, studies on outliers have been an important research topic for decades.

Since 18^{th} century, there have been many studies about the outlier detection. Bernoulli (1777) pointed out the practice of deleting the outliers about 200 years ago. Deletion of outliers is not a proper solution to handle the outliers but

this remained a common practice in past. To address the problem of outliers in the data, the first statistical technique was developed in 1850 (Beckman and Cook, 1983).

To detect the presence of outlier in data many methods have been developed. Some are graphical which may intuitively help, others are model-based. Methods can vary based on fields in which it's used. For instance, in regression, to check outliers or influential observations Cook's distance, Mahalanobis distance, DFFITS, DFBETAS, Studentized residuals and Dixon's Q test are widely used, for more details see Cook and Weisberg (1982).

1.1. Literature Survey on Outliers in Time Series

Time series observations may sometimes be affected by unusual events, disturbances, or errors that create spurious effects in the series and result in extraordinary patterns in the observations that are not in accord with most observations in the time series. Such unusual observations may be referred to as outliers (Box et al., 1994).

It is known that existing of outliers in time series can cause misidentification of the model, the biased parameter estimation and, consequently, poor forecasting. These reasons gave importance to outlier detection in time series. Thus studies on outlier detection in time series analysis have increased in the short term.

Fox (1972) made the first study on outlier detection in time series. He defined two types of outliers which possibly occur in the autoregressive (AR) time series as additive outlier (AO) and innovation outlier (IO), and developed the likelihood ratio test to determine outliers in AR models. Chang (1982) and Chang and Tiao (1983) extended Fox's test to ARIMA models and proposed an iterative procedure to detect multiple outliers (Tsay, 1988). After two outlier types presented by Fox (1972), Tsay (1988) introduced the new outlier types as level

shifts (LS) and temporary change (TC) and also focused on the structure changes and presented variance change.

Bruce and Martin (1989) proposed diagnostics for ARIMA model fitting for time series formed by deleting observations from the data and measuring the change in the estimates of the parameters. Abraham and Chuang (1989) have proposed tests, which are based on the deletion of the observation and similar with the one that Draper and John (1981) suggested for regression situation, to detect influential observations and outliers in time series. Pena (1990) investigated influential observations on coefficients in ARIMA models and introduced influence statistics based on Mahalanobis distance to measure the effect of two types of outliers on model parameters. Hui and Lee (1992) introduced a leave-one-out procedure, which has been widely used in regression analysis, to identify influential points. Lin and Lee (2001) adapted conditional Cook's distance, which is suggested by Lawrance and used in regression analysis, to the autoregressive models.

Kaiser and Maravall (1999) introduced a new outlier type called seasonal level shift (SLS).

Tsay, Pena and Pankratz (2000) extended the outlier types used in the univariate time series analysis to multiple cases and presented an iterative method which is based on two test statistics for estimating multivariate outliers.

Pena and Sanchez (2003) proposed a procedure that keeps the powerful features of previous methods but improves the initial parameter estimate, avoids the confusion between innovative outliers and level shift.

For other important studies on outliers in time series, see Chang et al. (1988), Abraham and Yatawara (1988), Hui and Lee (1992), Chen and Liu (1993), Ljung (1993), and Lu et al. (2012).

Now that there exist lots of outlier detection methods according to outlier types on variety time series models, in this study, outlier detection methods of AO, IO, TC and LC on ARIMA models are considered.

2. TIME SERIES

A time series can be defined as sequential data points that obtained with regular time intervals. A time series can be denoted as $Z_t, \mathrm{t} = 1, 2, ..., \mathrm{n}$ where n is length of series. Time series analysis involves statistical analysis of data which is obtained regularly over time, modeling and make prediction or forecasting by relevant model.

Observations that collected equally over time have considerably common place in today world. In the work area, we observe weekly interest rates, daily closing stock prices, monthly price indices, yearly sales figures, and so on. In meteorology, we observe daily high and low temperatures, annual precipitation and drought indices, and hourly wind speeds. In agriculture, we record annual figures for crop and livestock production, soil erosion, and export sales. In the biological sciences, we observe the electrical activity of the heart at millisecond intervals. In ecology, we record the abundance of an animal species. It can be say that areas in which time series are studied is almost endless (Cryer and Chan, 2008). There are many data sets that appear as time series: a monthly sequence of the quantity of goods delivered from a store, a weekly series of the number of highway accidents, hourly observations made on the yield of a chemical process, and so on. An intrinsic feature of a time series is that, typically, adjacent observations are dependent. Time series analysis is concerned with techniques for the analysis of this dependence. (Box et al., 1994)

2.1. Components of a Time Series

Time series' components consist of *trend*, *seasonality*, *cyclic variations* and *random* or *irregular movements*. The trend displays the general inclination of the observations to increase or decrease during a long period of time. The seasonality is short term movements occurring in data due to seasonal factors. The

cyclic variations can be defined as fluctuation which occurs in a time series. Lastly, irregular movements are unpredictable and abrupt changes, which occurs in a time series and are unlikely to be repeated.

One example of a time series with decreasing trend, which is shown in Figure 2.1., is the time series of US Treasury bill contracts on the Chicago market for 100 consecutive trading days in 1981 (Strickland, 2015).

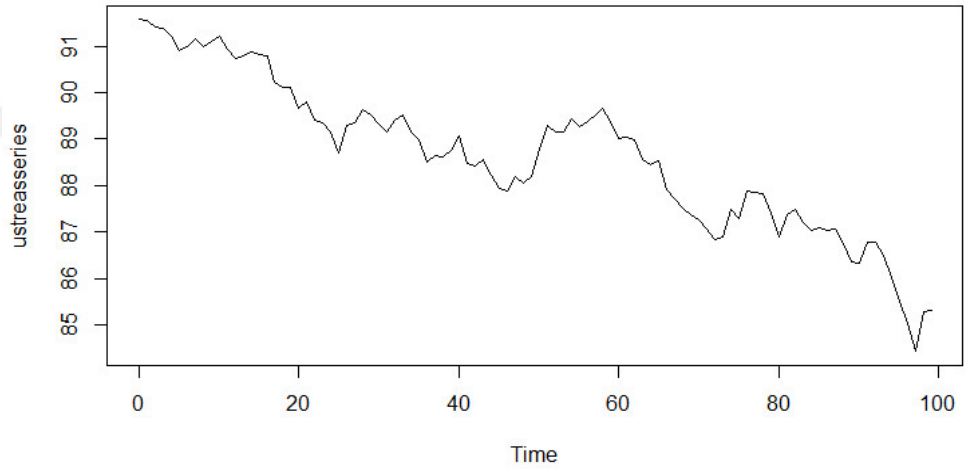


Figure 2.1. US Treasury bill contracts data plot

An example of a time series with seasonality is the time series of the seasonal US Birth data from 1946 to 1977, as shown in Figure 2.2. (Strickland, 2015).

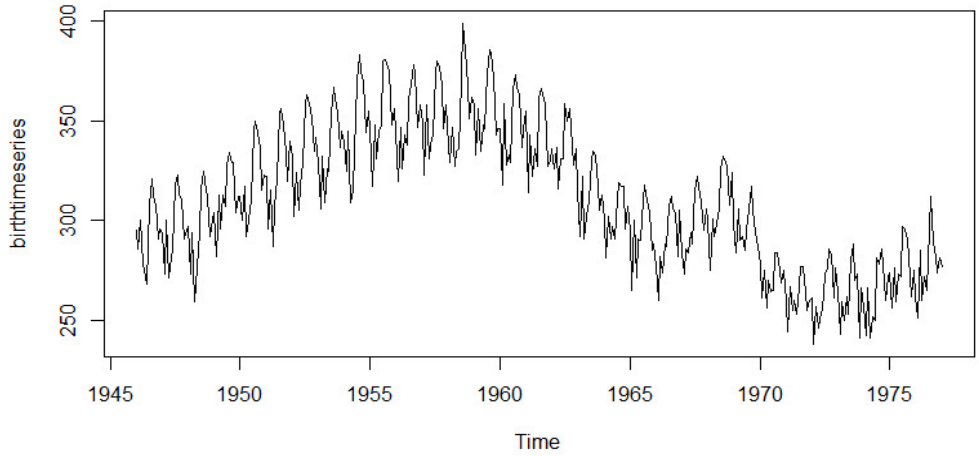


Figure 2.2. US Birth data plot

An example of a time series with cyclic variations is pork retailer prices in Germany from April 1995 to April 2012, as shown in Figure 2.3. (Strickland, 2015).

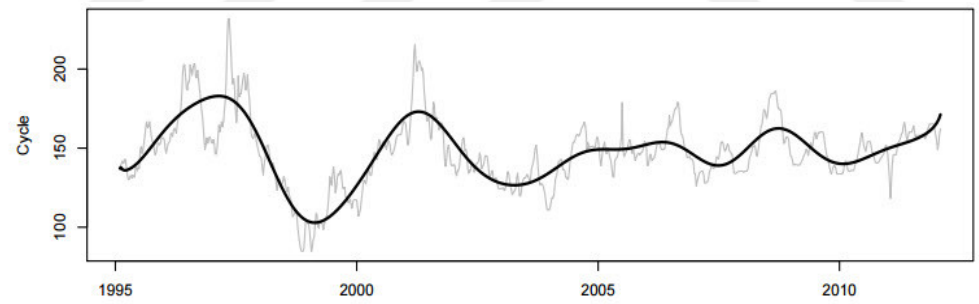


Figure 2.3. Pork retailer prices in Germany

2.2. Some Simple Operators

There are some operators employed extensively in time series analysis such as backward shift operator B, forward shift operator F and backward difference operator ∇. Briefing introductory explanations for these operators enable to understand the following sections more easily.

The backward shift operator also known as lag operator is defined by $B^m Z_t = Z_{t-m}$.

The forward shift operator is inverse of backward shift operator such that $F = B^{-1}$ and is defined by $F^m Z_t = Z_{t+m}$.

The backward difference operator is defined by $\nabla^m Z_t = Z_t - Z_{t-m}$ and can be written in terms of B such that $\nabla^m Z_t = (1 - \mathrm{B}^m)\, \mathrm{Z}_t$.

2.3. Autocovariance and Autocorrelation Functions

A time series can be denoted as $Z_t, \mathrm{t} = 1, 2, ..., \mathrm{n}$ where n is length of series. The covariance between Z_t and Z_{t+k}, separated by k interval of time, is called the *autocovariance* at lag k and is defined by

$$\gamma_k = Cov(Z_t, Z_{t+k}) = \mathrm{E}\left[(Z_t - \mu)(Z_{t+k} - \mu)\right].$$

Similarly, the *autocorrelation* at lag k is (Box et al., 1994)

$$\rho_k = \frac{\mathrm{E}\left[(Z_t - \mu)(Z_{t+k} - \mu)\right]}{\sqrt{\mathrm{E}\left[(Z_t - \mu)^2\right]\mathrm{E}\left[(Z_{t+k} - \mu)^2\right]}} = \frac{Cov(Z_t, Z_{t+k})}{\sigma_{Z_t}\sigma_{Z_{t+k}}}.$$

2.4. Concept of Stationary

Concept of stationary is vital in time series analysis because many statistical models and methods used in time series analysis assume that series is stationary.

A valuable class of stochastic models to describe time series, which has received a wide range of attention, comprises what are called stationary models, which assume that the process remains in equilibrium about a constant mean level (Box et al., 1994).

To make statistical inferences about the structure of a stochastic process on the basis of an observed record of that process, we must usually make some simplifying (and presumably reasonable) assumptions about that structure. The most important such assumption for time series is that of stationarity. The basic idea of stationarity is that the probability laws that govern the behavior of the process do not change over time. In a sense, the process is in statistical equilibrium. Specifically, a process (Z_t) is said to be *strictly stationary* if the joint distribution of $Z_{t_1}, Z_{t_2}, ..., Z_{t_n}$ is the same as the joint distribution of $Z_{t_{1-k}}, Z_{t_{2-k}}, ..., Z_{t_n-k}$ for all choices of time points $t_1, t_2, ..., t_n$ and all choices of time at lag k (Cryer and Chan, 2008).

Thus, when $n = 1$ the (univariate) distribution of Z_t is the same as that of Z_{t-k} for all t and k; in other words, the Z's are (marginally) identically distributed. It then follows that $\mathrm{E}(\mathrm{Z}_t) = \mathrm{E}(\mathrm{Z}_{t-k})$ for all t and k so that the mean function is constant for all time. Additionally, $Var(\mathrm{Z}_t) = Var(\mathrm{Z}_{t-k})$ for all t and k so that the variance is also constant over time (Cryer and Chan, 2008).

Setting n = 2 in the stationarity definition we see that the bivariate distribution of Z_t and Z_s must be the same as that of Z_{t-k} and Z_{s-k} from which it follows that $Cov(\mathrm{Z}_t, \mathrm{Z}_s) = Cov(\mathrm{Z}_{t-k}, \mathrm{Z}_{s-k})$ for all *t*, *s*, and *k*. Putting *k* = *s* and then *k*=*t*, we obtain $\gamma_{t,s} = Cov(\mathrm{Z}_{t-s}, \mathrm{Z}_0) = Cov(\mathrm{Z}_0, \mathrm{Z}_{s-t}) = Cov(\mathrm{Z}_0, \mathrm{Z}_{|t-s|}) = \gamma_{0,|t-s|}$; that is, the covariance between Z_t and Z_s depends on time only through the time different $|t-s|$ and not otherwise on the actual times t and s. If a process is strictly stationary and has finite variance, then the covariance function must depend only on the time lag. A stochastic process (Z_t) is said to be *weakly* (or second-order)

stationary if the mean function is constant over time, and $\gamma_{t,t-k} = \gamma_{0,k}$ for all time t and lag k (Cryer and Chan, 2008).

If neither the mean μ_t nor the autocovariances $\gamma_{j,t}$ depend on the time t, then the process for Z_t is said to be covariance-stationary or *weakly stationary* (Hamilton, 1994):

$$E(Z_t) = \mu \qquad \text{for all } t$$
$$\mathrm{E}\left[(Z_t - \mu)(Z_{t+j} - \mu)\right] = \gamma_j \quad \text{for all } t \text{ and any } j.$$

A different concept is that of *strict stationarity*. A process is said to be strictly stationary if, for any values of $j_1, j_2, \ldots, \mathrm{j}_n$, the joint distribution of $(Z_t, Z_{t+j_1}, Z_{t+j_2}, ..., Z_{t+j_n})$ depends only on the intervals separating the times $(j_1, j_2, \ldots, \mathrm{j}_n)$ and not on the time itself (t). Notice that if a process is strictly stationary with finite second moments, then it must be covariance-stationary. However, it is possible to imagine a process that is covariance-stationary but not strictly stationary (Hamilton, 1994).

A time series is *white noise* if the variables are independent and identically distributed with a mean of zero. This means that all variables have the same variance σ^2 and each value has a zero correlation with all other values in the series. That is, assume ε_t is white noise, then

$$E(\varepsilon_t) = 0,$$
$$E(\varepsilon_t^2) = \sigma^2,$$
$$E(\varepsilon_t \varepsilon_j) = 0 \quad \text{for } t \neq j.$$

This is simply denoted as $\varepsilon_t \sim WN(0, \sigma^2)$ and called as white noise.

In practice, many real time series do not ensure stationarity and can have components of time series since they are not in statistical equilibrium. However, it is possible to transform non-stationary series into stationary series by simple techniques such as differencing.

In literature, there exist many both graphical and statistical methods about whether or not a time series is stationary. ACF graph of a series may help to comment about the stationarity of the series. Non-stationary series have gradually decreasing autocorrelations that can be a function of time, whereas stationary series have stable but rapidly decreasing autocorrelations.

Here, significance test for autocorrelation coefficient and unit root tests, which are effective and widely used to distinguish the stationarity of the series, are considered.

2.4.1. Significance Test for Autocorrelation Coefficient

For a stationary process, the variance $\sigma_Z^2 = \gamma_0$ is the same at time $t+k$ as at time t. Thus the autocorrelation at lag k; that is, the correlation between Z_t and Z_{t+k}, is

$$\rho_k = \frac{Cov(Z_t, Z_{t+k})}{\sigma_{Z_t}\sigma_{Z_{t+k}}} = \frac{Cov(Z_t, Z_{t+k})}{\sigma_{Z_t}^2} = \frac{\gamma_k}{\gamma_0}$$

which implies that $\rho_k = 1$ for k=0. To predict ρ_k from a given sample with n size,

$$\hat{\gamma}_k = \frac{\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{n}, \text{ and } \hat{\gamma}_0 = \frac{\sum_{t=1}^{n-k}(Z_t - \bar{Z})^2}{n}$$

Hence,

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}\left(Z_t - \bar{Z}\right)\left(Z_{t+k} - \bar{Z}\right)}{\sum_{t=1}^{n-k}\left(Z_t - \bar{Z}\right)^2} = ACF_{(k)}$$

To assess whether or not series is stationary, hypothesis of significance test for autocorrelation coefficient is as below. For k=1,2,…,n

$$\mathrm{H}_0 : \rho_k = 0$$
$$\mathrm{H}_1 : \rho_k \neq 0$$

Test statistic corresponding to this hypothesis is

$$t_{ACF_{(k)}} = \frac{ACF_{(k)}}{Se\left(ACF_{(k)}\right)}$$

where $Se\left(ACF_{(k)}\right)$ is the standard error of the autocorrelation and defined as

$$Se\left(ACF_{(k)}\right) \cong \frac{1}{\sqrt{n}}\left(1 + 2\sum_{k=1}^{q}\rho_k^2\right)$$

where q is such that all the autocorrelations are zero for $k>q$. For large n, Box et al. (1994) suggested that the standard error of ACF approximately equals to $1/\sqrt{n}$.

The table of critical values to compare with test statistic is Student's t-table which is widely known.

If the test statistic is greater than the critical value, then the null hypothesis is rejected which means time series is non-stationary. If the test statistic is not as extreme as the critical value, then the null hypothesis is not rejected which means time series is stationary.

To test k-autocorrelation coefficient at the same time, Q statistic of Ljung and Box (1978), which is closely connected to Box-Pierce test, is popular and effective one. Hypothesis for this test is

$$\mathrm{H}_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$$
$$\mathrm{H}_1 : \exists \rho_i \neq 0, i = 1, 2, ..., \mathrm{k}$$

And test statistic is

$$Q = n(\mathrm{n}+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{(n-k)} \ ,$$

where h is the maximum lag being considered (usually $h \cong 20$ is selected) and n is the number of observations. Under H_0 hypothesis, the statistic Q asymptotically follows $\chi^2_{(h-m)}$, which is chi-squared distribution with (h-m) degrees of freedom and m is the number of parameters in the model, which has been fitted to the data. For significance level α, the critical region for the rejection of the hypothesis, which means time series is non-stationary, is: $Q > \chi^2_{1-\alpha,h-m}$ where $\chi^2_{1-\alpha,h-m}$ is the $1-\alpha$ quantile of the chi-squared distribution with (h-m) degrees of freedom.

2.4.2. Unit Root Tests

Another way of determining whether or not time series is stationary is unit root test. If time series does not have unit root then time series is stationary, in contrast, time series is not stationary.

Although there are lots of tests to determine the existence of unit root in time series such as Dickey-Fuller, Phillips-Person, Schmidt-Philips, KPSS, and others, just the most popular ones among them, which are Dickey-Fuller's tests, are considered to detect the existence of unit root in this study.

2.4.2.1. The Dickey-Fuller Test

To explain how to imply the test, consider Z_t which follows $\mathrm{AR}(1)$, which indicates an autoregressive model of order 1 process

$$Z_t = \phi Z_{t-1} + \varepsilon_t \tag{0.1}$$

where ε_t is a sequence of independent random variables with mean zero and variance σ^2 and shortly denoted as $\varepsilon \sim WN(0,\sigma^2)$ and ϕ is the parameter of the model. The time series Z_t converges (as $t \to \infty$) to a stationary time series if $|\phi| < 1$. If $|\phi|$=1, the time series is not stationary and the variance of Z_t is $t\sigma^2$. The time series with $\phi = 1$ is sometimes called a *random walk*. If $|\phi| > 1$, the time series is not stationary and the variance of the time series grows exponentially as t increases (Dickey and Fuller, 1979).

To test null hypothesis $\mathrm{H}_0 : \phi = 1$, which means the time series has unit root that is not stationary, test statistic given by Dickey-Fuller (1979) is as below

$$\hat{\tau} = \frac{\left(\hat{\phi} - 1\right)\left(\sum_{t=2}^{n} Z_{t-1}^2\right)^{1/2}}{S_e}$$

where

$$S_e^2 = \frac{\sum_{t=2}^{n} (Z_t - \hat{\phi} Z_{t-1})^2}{n-2}$$

and $\hat{\phi}$ is, when given n observations $Z_1, \ldots, Z_n$, the least squares estimator of ϕ is given as

$$\hat{\phi} = \frac{\sum_{t=1}^{n} Z_t Z_{t-1}}{\sum_{t=1}^{n} Z_{t-1}^2}$$

The table of critical values of the test statistic is calculated according to three different model types, which are constant and trend free model, including constant and including both constant and trend, and can be found in many time series sources (see Fuller, 1996 or Hamilton, 1994). Consequently, if the test statistic is more negative than the critical value, then the null hypothesis is rejected which means no unit root is present in time series; in other words, we can say time series is stationary.

There is also an extension of the Dickey–Fuller (DF) test called the augmented Dickey–Fuller test (ADF), which removes all the structural effects (autocorrelation) in the time series and then tests using the same procedure. In practice, ADF is more common because it can handle more complex models than DF and is also more powerful.

2.4.2.2. Augmented Dickey-Fuller Test

This test is an extension of the Dickey–Fuller (DF) test showed previously. To show how to imply ADF test, if we subtract the term Z_{t-1} from both sides of Eq. (2.1), we have

$$\begin{aligned} \nabla Z_t = Z_t - Z_{t-1} &= \phi Z_{t-1} - Z_{t-1} + \varepsilon_t \\ &= (\phi - 1) Z_{t-1} + \varepsilon_t \end{aligned} \tag{0.2}$$

Now that if we denote the term $\phi - 1$ as ψ, then testing whether ψ is equal to zero in Eq. (2.2) will be same as testing whether ϕ is equal to 1 in Eq. (2.1). Note that time series given in Eq. (2.1) is not stationary under the null hypothesis that ϕ is equal to 1; similarly, time series given in Eq. (2.2) is not stationary under the null hypothesis that ψ is equal to zero.

The testing procedure for the ADF test is the same as for the Dickey–Fuller test, but it is applied to the model

$$\begin{aligned} \nabla Z_t &= c + c_1 t + \psi Z_{t-1} + \phi_1 \nabla Z_{t-1} + \cdots + \phi_{p-1} \nabla Z_{t-p+1} + \varepsilon_t \\ &= c + c_1 t + \psi Z_{t-1} + \sum_{i=1}^{p-1} \phi_i \nabla Z_{t-i} + \varepsilon \end{aligned} \tag{0.3}$$

where c is constant, c_1 is parameter and others are as defined former.

To test null hypothesis $H_0 : \psi = 0$, which means the time series has unit root; that is not stationary, test statistic given by Dickey-Fuller (1981) is as below

$$\hat{\tau} = \frac{\hat{\psi}}{Se(\hat{\psi})}$$

where $\hat{\psi}$ is the least squares estimator of ψ in Eq. (2.3) and $Se(\hat{\psi})$ is the standard error of $\hat{\psi}$. The critical values corresponding to test statistic is same as the Dickey-Fuller test. Similarly as in the Dickey-Fuller test decision, if the test statistic is more negative than the critical value, then the null hypothesis is rejected which means that no unit root is present, in other words, time series is stationary.

2.5. Time Series Models

Time series models can be classified as linear and nonlinear models, and each of them can be separated two types according to stationarity of the model.

A linear time series is one where, for each observation Z_t , that observation can be viewed as a linear combination of past or future values or differences.

Nonlinear time series are generated by nonlinear dynamic equations. They display features that cannot be modeled by linear processes: time-changing variance, asymmetric cycles, higher-moment structures, thresholds and breaks.

Time series which have the representation $Z_t = \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k}$ where ε_t are identically independent distributed random variables. Such models are extremely useful, because they are designed to model the autocovariance structure and are straightforward to use for forecasting. These are some of the reasons that they are widely used in several applications. A typical realization from a linear time series will be quite regular with no sudden bursts or jumps. This is due to the linearity of the system. However, if one looks at financial data, for example, there are sudden bursts in volatility (variation) and extreme values, which calm down after a while. It is not possible to model such behaviour well with a linear time series. In order to capture ‘nonlinear behaviour several nonlinear models have been proposed. The models typically consist of products of random variables, which make possible the sudden irratic bursts seen in the data. Over the past 30 years, there has been a lot

research into nonlinear time series models. Probably one of the first nonlinear models proposed for time series analysis is the bilinear model, this model is used extensively in signal processing and engineering. A popular model for modelling financial data is (G)ARCH-family of models. Other popular models are random autoregressive coefficient models and threshold models, to name but a few (see, for example, Subba Rao (1977), Granger and Andersen (1978), Nicholls and Quinn (1982), Engle (1982), Subba Rao and Gabr (1984), Bollerslev (1986), Terdik (1999), Fan and Yao (2003), Straumann (2005) and Douc et al. (2014) (Subba R.S., 2008).

In this study, linear time series according to stationarity are considered.

2.5.1. Stationary Time Series Models

Linear time series models are designed to model the covariance structure in the time series. There are two popular subgroups of linear time model: the autoregressive and the moving average models, which can be combined to make the autoregressive moving average models (Subba Rao, S., 2018). These models have assumed great importance in modelling real-world processes.

2.5.1.1. Autoregressive Process

The models expressing the observation values in any period of a time series as the linear combination of the observation values of a certain number of previous periods of the same series and with error term are called autoregressive (AR) models. As its name signifies, AR processes are regressions on themselves. An $\mathrm{AR}(p)$ model is an autoregressive model where specific lagged values of Z_t are used as predictor variables. Lags are where results from one time period affect following periods.

2.5.1.1.(1). A pth-order autoregressive process

The value "p" is called the order of process. Model of p th-order autoregressive process, succinctly $\mathrm{AR}(p)$, is as below

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \varepsilon_t \tag{0.4}$$

where $\varepsilon \sim WN(0, \sigma^2)$, c is a constant term and $\phi_1, \phi_2, ..., \phi_p$ are unknown autoregressive parameters. This is like a multiple regression but with *lagged values* of Z_t as predictors.

The reason for autoregressive process name is that a linear model

$$Z_t = c + \phi_1 x_1 + \phi_2 x_2 + \cdots + \phi_p x_p + \varepsilon_t$$

relating a "dependent" variable Z to a set of "independent" variables (predictors) $x_1, x_1, ..., x_p$, plus constant c and error term ε_t, is often referred to as a *regression model*, and Z is said to be "regressed" on $x_1, x_1, ..., x_p$. In Eq. (2.4) the variable Z is regressed on previous values itself; hence the model is *autoregressive* (Box et al., 1994).

The model given in Eq. (2.4) can be rewritten (assumed no constant, this also means that the term constant has been suppressed by assuming that the variable has been measured as deviations from its mean) by using backshift operator (B) as

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) Z_t = \varepsilon_t \tag{0.5}$$

or succinctly, using autoregressive operator $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ which denotes the left side of Z_t in Eq. (2.5) as

$$\phi(B) Z_t = \varepsilon_t . \tag{0.6}$$

Since Eq. (2.6) implies that

$$Z_t = \phi^{-1}(B)\varepsilon_t$$

The autoregressive process can be taught of as the Z_t output from a linear filter with transfer function $\phi^{-1}(\mathrm{B})$ when input is white noise ε_t (Box et al., 1994).

Autoregressive process can be stationary or nonstationary. Eq. (2.6) must satisfy certain conditions for the process to be stationary. The necessary requirement for stationarity is that the autoregressive operator, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, considered as a polynomial in B of degree p, must have all roots of $\phi(B) = 0$ greater than 1 in absolute value; that is, all roots must lie outside the unit circle.

The mean of an $\mathrm{AR}(p)$ process $E(Z_t) = \mu$ is

$$E(Z_t) = E\left(c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \varepsilon_t\right)$$
$$\mu = c / \left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right). \tag{0.7}$$

The mean of $\mathrm{AR}(p)$ does not depend on time. To obtain the variance of an $\mathrm{AR}(p)$ process, Eq. (2.4) can be written by using Eq. (2.7) as below.

$$Z_t - \mu = \phi_1 (Z_{t-1} - \mu) + \phi_2 (Z_{t-2} - \mu) + \cdots + \phi_p (Z_{t-p} - \mu) + \varepsilon_t$$

Now multiplying both sides of this equation with $Z_{t-k} - \mu$ and then taking expectations, we get

$$(Z_t-\mu)(Z_{t-k}-\mu)=\phi_1(Z_{t-1}-\mu)(Z_{t-k}-\mu)+\phi_2(Z_{t-2}-\mu)(Z_{t-k}-\mu)+\cdots$$
$$+\phi_p(Z_{t-p}-\mu)(Z_{t-k}-\mu)+(Z_{t-k}-\mu)\varepsilon_t$$
$$E\left[(Z_t-\mu)(Z_{t-k}-\mu)\right]=\phi_1E\left[(Z_{t-1}-\mu)(Z_{t-k}-\mu)\right]+\phi_2E\left[(Z_{t-2}-\mu)(Z_{t-k}-\mu)\right]+$$
$$\cdots+\phi_pE\left[(Z_{t-p}-\mu)(Z_{t-k}-\mu)\right]+E\left[(Z_{t-k}-\mu)\varepsilon_t\right]$$

Since $E\left[(Z_{t-k}-\mu)\varepsilon_t\right]$ is equal to zero when $k>0$, we obtain

$$\gamma_k=\phi_1\gamma_{k-1}+\phi_2\gamma_{k-2}+\cdots+\phi_p\gamma_{k-p}\quad,k>0 \tag{0.8}$$

Now when $k=0$, the term $E\left[(Z_{t-k}-\mu)\varepsilon_t\right]$ will be

$$\begin{aligned}E\left[(Z_t-\mu)\varepsilon_t\right]&=E\left[(\phi_1Z_{t-1}+\phi_2Z_{t-2}+\cdots+\phi_pZ_{t-p}+\varepsilon_t)\varepsilon_t\right]\\&=E\left[\phi_1Z_{t-1}\varepsilon_t+\phi_2Z_{t-2}\varepsilon_t+\cdots+\phi_pZ_{t-p}\varepsilon_t+\varepsilon_t^2\right]\\&=\phi_1E(Z_{t-1}\varepsilon_t)+\phi_2E(Z_{t-2}\varepsilon_t)+\cdots+\phi_pE(Z_{t-p}\varepsilon)_t+E(\varepsilon_t^2)\\&=E(\varepsilon_t^2)\\&=\sigma_\varepsilon^2.\end{aligned}$$

Then the variance of $\mathrm{AR}(p)$ equals to

$$\gamma_0=\phi_1\gamma_{-1}+\phi_2\gamma_{-2}+\cdots+\phi_p\gamma_{-p}+\sigma_\varepsilon^2$$

and also can be written by dividing γ_0 and substituting $\gamma_k=\gamma_{-k}$,

$$\gamma_0=\sigma_Z^2=\sigma_\varepsilon^2/(1-\phi_1\rho_1+\phi_2\rho_2+\cdots+\phi_p\rho_p).$$

The autocovariance of $\mathrm{AR}(p)$ does not depend on time. Now since we have both variance and autocovariance of $\mathrm{AR}(p)$, autocorrelation of the process can be obtained by dividing Eq. (2.8) with γ_0 as

$$\rho_k=\frac{\gamma_k}{\gamma_0}=\phi_1\rho_{k-1}+\phi_2\rho_{k-2}+\cdots+\phi_p\rho_{k-p}\quad,k>0.$$

In general, the autocorrelation function of a stationary autoregressive process consists of a mixture of damped exponentials and damped sine waves.

In practice, $AR(1)$ and $AR(2)$ models are significantly used. Now we will focus on these models and their structure properties.

2.5.1.1.(2). A first-order autoregressive process

An $AR(1)$ is an expression relating a variable Z_t to its first lag value Z_{t-1} as

$$Z_t = c + \phi_1 Z_{t-1} + \varepsilon_t \tag{0.9}$$

where $\varepsilon \sim WN(0, \sigma^2)$, c is a constant term, and ϕ_1 is unknown autoregressive parameter.

In Eq. (2.9), ϕ_1 must satisfy the condition $-1 < \phi_1 < 1$ for process to be stationary. Besides the parameter's satisfying condition $-1 < \phi_1 < 1$, if $AR(1)$ process is said to be stationary, then the mean variance and autocovariance of the process do not change depending on time, that is

$E(Z_t) = E(Z_{t-1}) = \mu$,

$Var(Z_t) = Var(Z_{t-1}) = \sigma_Z^2$,

$\mathrm{E}\left[(Z_t - \mu)(Z_{t+k} - \mu)\right] = \gamma_k$ for all t and any k.

To examine the stationarity of the process, it is necessary to check whether these conditions are met. The mean of a $AR(1)$ process is

$$E(Z_t) = \mu = c/(1-\phi_1).$$

The mean of $AR(1)$ does not depend on time. The variance of a $AR(1)$ process is

$$\sigma_Z^2 = \gamma_0 = Var(c + \phi_1 Z_{t-1} + \varepsilon)$$

as

$$\sigma_Z^2 = \sigma_\varepsilon^2 / (1 - \phi_1^2)$$

The autocovariance of an $AR(1)$ process can be obtained from

$$\gamma_k = Cov(Z_t, Z_{t-k})$$
$$= E\left[(Z_t - \mu)(Z_{t-k} - \mu)\right]$$

By considering

$$Z_t = \mu + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} + \cdots + \phi_1^k\varepsilon_{t-k} + \phi_1^{k+1}\varepsilon_{t-k-1} + \phi_1^{k+2}\varepsilon_{t-k-2} + \cdots$$
$$Z_{t-k} = \mu + \varepsilon_{t-k} + \phi_1\varepsilon_{t-k-1} + \phi_1^2\varepsilon_{t-k-2} + \cdots$$

as

$$\gamma_k = \left[\phi_1^k + \phi_1^{k+2} + \phi_1^{k+4} + \cdots\right]\sigma_\varepsilon^2$$
$$= \left[\phi_1^k\left(1 + \phi_1^2 + \phi_1^4 + \cdots\right)\right]\sigma_\varepsilon^2 \quad \text{(Hamilton, 1994)}$$
$$= \left[\phi_1^k / \left(1 - \phi_1^2\right)\right]\sigma_\varepsilon^2$$

Autocorrelation of an $AR(1)$ process is obtained as

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\left[\phi_1^k / \left(1 - \phi_1^2\right)\right]\sigma_\varepsilon^2}{\sigma_\varepsilon^2 / \left(1 - \phi_1^2\right)} = \phi_1^k, \ \mathrm{k} \geq 0.$$

We simulated n=50 observations, which is shown in Figure 2.4., from the AR(1) model $Z_t = \phi Z_{t-1} + \varepsilon_t$ by using RStudio with different phi values respectively 0.6 and -0.4, and $\sigma_\varepsilon^2 = 1$.

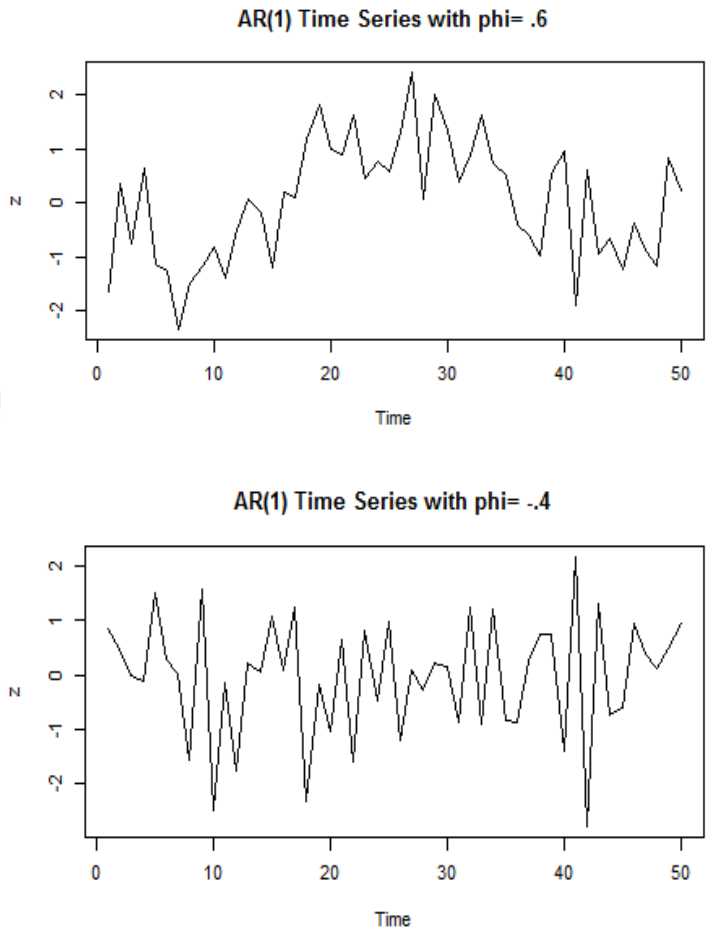


Figure 2.4. Plots of simulated AR(1) series with different phi values

A particular type of $\mathrm{AR}(1)$ is the situation $\phi_1 = 1$, which means time series is non-stationary, and the process is called as ***random walk*** which is defined as

$$Z_t = Z_{t-1} + \varepsilon_t \tag{0.10}$$

where $\varepsilon \sim WN(0, \sigma^2)$. It can easily be seen by looking at the variance that the random walk process is not stationary. To find the variance of random walk process, first consider that Eq. (2.10) can be rewritten as

$$\begin{aligned} Z_t &= Z_{t-1} + \varepsilon_t \\ &= \left(Z_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t \\ &= \left(Z_{t-3} + \varepsilon_{t-2}\right) + \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ &= \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t . \end{aligned}$$

Now the variance of random walk process is

$$\begin{aligned} Var(Z_t) &= Var\left(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t\right) \\ &= Var\left(\varepsilon_1\right) + Var\left(\varepsilon_2\right) + \cdots + Var\left(\varepsilon_t\right) \quad (\varepsilon_t\text{'s are independent}) \\ &= \sigma_\varepsilon^2 + \sigma_\varepsilon^2 + \cdots + \sigma_\varepsilon^2 \\ &= t\sigma_\varepsilon^2 . \end{aligned}$$

As time increases, so does the variance, which shows that process is nonstationary.

In Figure 2.5 we simulated n=50 observations from the random walk model $Z_t = Z_{t-1} + \varepsilon_t$, where phi is 1 and $\sigma_\varepsilon^2 = 1$ by using RStudio.

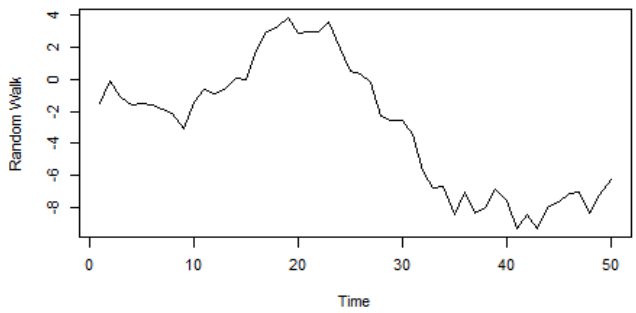


Figure 2.5. Plot of simulated random walk series

2.5.1.1.(3). A second-order autoregressive process

Model of second-order autoregressive process, succinctly $\mathrm{AR}(2)$, is as below

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \varepsilon_t \tag{0.11}$$

where $\varepsilon \sim WN(0,\sigma^2)$, c is a constant term, and ϕ_1, ϕ_2 are unknown autoregressive parameters. For stationarity, the parameters ϕ_1 and ϕ_1 must satisfy the conditions

$$\begin{aligned} &\phi_2 + \phi_1 < 1 \\ &\phi_2 - \phi_1 < 1 \\ &-1 < \phi_2 < 1. \end{aligned} \tag{0.12}$$

Besides parameters' satisfying conditions given in Eq. (2.12), if $\mathrm{AR}(2)$ process is said to be stationary, then the mean, variance and autocovariance of the process do not change depending on time. The mean of $\mathrm{AR}(2)$ process $E(Z_t) = \mu$ can be obtained from

$$E(Z_t) = E(c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \varepsilon_t)$$

as

$$\mu = c / (1 - \phi_1 - \phi_2)$$

To find the variance of an $\mathrm{AR}(2)$ process, Eq. (2.11) can be rewritten as

$$\begin{aligned} Z_t &= \mu(1 - \phi_1 - \phi_2) + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \varepsilon_t \\ Z_t - \mu &= \phi_1 (Z_{t-1} - \mu) + \phi_2 (Z_{t-2} - \mu) + \varepsilon_t. \end{aligned} \tag{0.13}$$

Now multiplying both sides of Eq. (2.13) by $(Z_t - \mu)$ and taking expectations, we get

$$E(Z_t-\mu)^2 = \phi_1 E\left[(Z_t-\mu)(Z_{t-1}-\mu)\right] + \phi_2 E\left[(Z_t-\mu)(Z_{t-2}-\mu)\right] + E\left[(Z_t-\mu)\varepsilon_t\right]$$
$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + E\left[(Z_t-\mu)\varepsilon_t\right] \quad (0.14)$$

where

$$\begin{aligned} E\left[(Z_t-\mu)\varepsilon_t\right] &= E\left[\left[\phi_1(Z_{t-1}-\mu)+\phi_2(Z_{t-2}-\mu)+\varepsilon_t\right]\varepsilon_t\right] \\ &= \phi_1 E\left[(Z_{t-1}-\mu)\varepsilon_t\right] + \phi_2 E\left[(Z_{t-2}-\mu)\varepsilon_t\right] + E\left[\varepsilon_t^2\right] \\ &= \phi_1.0 + \phi_2.0 + \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \end{aligned} \quad (0.15)$$

Substituting (2.15) into (2.14) gives

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma_\varepsilon^2 \quad (0.16)$$

or similarly

$$\gamma_0 = \phi_1\rho_1\gamma_0 + \phi_2\rho_2\gamma_0 + \sigma_\varepsilon^2. \quad (0.17)$$

The variance of $AR(2)$ does not depend on time. To obtain the autocovariance of $AR(2)$ process, now multiply both sides of (2.13) by $(Z_{t-k} - \mu)$ and taking expectations produces

$$E\left[(Z_{t-k}-\mu)(Z_t-\mu)\right] = \phi_1 E\left[(Z_{t-k}-\mu)(Z_{t-1}-\mu)\right] + \phi_2 E\left[(Z_{t-k}-\mu)(Z_{t-2}-\mu)\right] + E\left[(Z_{t-k}-\mu)\varepsilon_t\right]$$

Since ε_t is independent of Z_{t-k}, we get

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} \quad , \; k > 0.$$

Autocorrelation of an $AR(2)$ process is

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi_1 \frac{\gamma_{k-1}}{\gamma_0} + \phi_2 \frac{\gamma_{k-2}}{\gamma_0}$$

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} \tag{0.18}$$

Autocorrelations belong to the first and second lag can be obtained from Eq. (2.18) as

$$\rho_1 = \frac{\phi_1}{1-\phi_2} \quad , \quad \rho_2 = \phi_2 + \frac{\phi_1^2}{1-\phi_2}. \tag{0.19}$$

The variance of $AR(2)$, in terms of ϕ_1 and ϕ_2, given in Eq. (2.16) can be obtained by substituting (2.19) into (2.17), as

$$\gamma_0 = \frac{(1-\phi_2)\sigma_\varepsilon^2}{(1+\phi_2)\left[(1-\phi_2)^2 - \phi_1^2\right]}.$$

In Figure 2.6, we simulated n=50 observations from the AR(2) model $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \varepsilon_t$, with phi parameters ϕ_1 respectively 0.5, 0.6 and ϕ_2 respectively 0.4,-0.3 and $\sigma_\varepsilon^2 = 1$ by using RStudio.

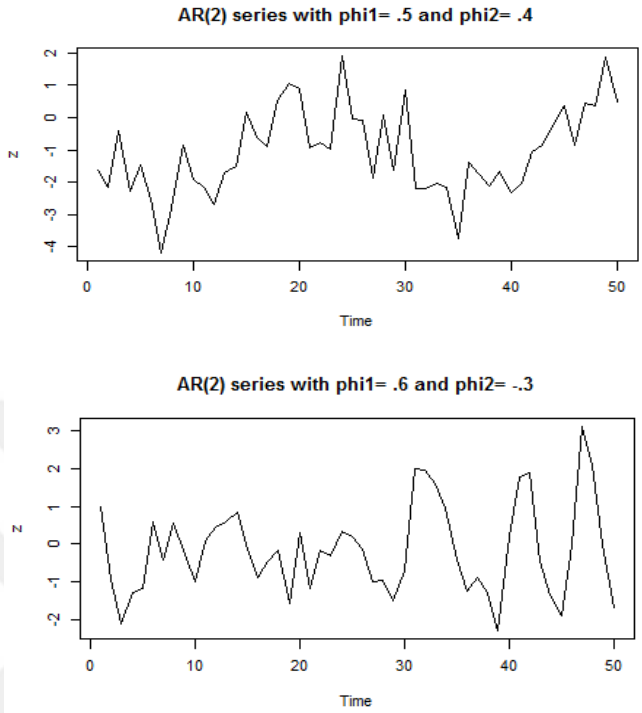


Figure 2.6. Plots of simulated AR(2) series

2.5.1.2. Moving Average Process

The models expressing the observation values in any period of a time series depending on error term in the same period and a certain number of previous periods' error term are called moving average (MA) models.

The term “moving average” comes from the fact that Z_t is constructed from a weighted sum, akin to an average, of the two most recent values of ε (Hamilton, 1994).

2.5.1.2.(1). A qth-order moving average process

Model of q th-order moving average process, succinctly $\mathrm{MA}(q)$, is as below

$$Z_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \qquad (0.20)$$

where $\varepsilon \sim WN(0,\sigma^2)$, c is a constant term (the mean of process) and $\theta_1, \theta_2, ..., \theta_q$ are unknown moving average parameters. This is like a multiple regression but with *lagged values* of ε_t as predictors.

A moving-average model given in Eq. (2.20) is conceptually a linear regression of the current value of the series against current and previous (observed) white noise error terms.

The mean of a $\mathrm{MA}(q)$ process $E(Z_t) = \mu$ is

$$E(Z_t) = E\left(c + \varepsilon + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}\right)$$
$$\mu = c\,.$$

The variance of a $\mathrm{MA}(q)$ process is

$$\gamma_0 = E(Z_t - \mu)^2 = E\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}\right)^2$$

Since the ε_t's are uncorrelated, expectation of product of different ε_t's lag is zero; that is, $E(\varepsilon_t\varepsilon_{t-k}) = 0 \;, k > 0$, then the variance is obtained as

$$\gamma_0 = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2.$$

The autocovariance of a $\mathrm{MA}(q)$ process is

$$\gamma_k = E\left[Z_t.Z_{t-k}\right]$$
$$= E\left[\theta_k\varepsilon_{t-k}^2 + \theta_{k+1}\theta_1\varepsilon_{t-k-1}^2 + \theta_{k+2}\theta_2\varepsilon_{t-k-2}^2 + \cdots + \theta_q\theta_{q-k}\varepsilon_{t-q}^2\right].$$

Terms involving ε_t's at different dates have been dropped because their product has expectation zero, and θ_0 is defined to be unity. For $k > q$, there are no

ε_t's with common dates in the definition of γ_k , and so the expectation is zero. Thus,

$$\gamma_k = \begin{cases} \left(\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 + \cdots + \theta_q\theta_{q-k}\right)\sigma_\varepsilon^2 , & \text{for } k = 1, 2, ..., q \\ 0 , & \text{for } k > q. \end{cases}$$

For any values of $\left(\theta_1, \theta_2, \ldots, \theta_q\right)$, the $MA(q)$ process is covariance stationary (Hamilton, 1994).

Contrary to AR model, the finite MA model is *always stationary*. Now since we have both variance and autocovariance of $\mathrm{MA}(q)$, the autocorrelation of the process can be obtained as

$$\rho_k = \begin{cases} \dfrac{\left(\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 + \cdots + \theta_q\theta_{q-k}\right)\sigma_\varepsilon^2}{\left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2} , & 0 < k \le q \\ 0 , & k > q \end{cases}$$

The autocorrelation function of a moving average process has a cutoff after lag q and this helps to determine the order of the process.

In practice, $\mathrm{MA}(1)$ and $\mathrm{MA}(2)$ models are significantly used. Now we will focus on these models and their structure properties.

2.5.1.2.(2). A first-order moving average process

A $\mathrm{MA}(1)$ is an expression relating a variable Z_t to current and one previous value of error term:

$$Z_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1}$$

where $\varepsilon \sim WN(0, \sigma^2)$, c is a constant term, and θ_1 is unknown parameter.

The mean of a $\mathrm{MA}(1)$ process is

$$E\left(Z_t\right)=\mu=c\,.$$

The variance of a $\mathrm{MA}(1)$ process is obtained by considering that the ε_t's are uncorrelated as

$$\begin{aligned}\gamma_0 &= E\left(Z_t-\mu\right)^2 = E\left(\varepsilon+\theta_1\varepsilon_{t-1}\right)^2\\ &= E\left(\varepsilon_t^2\right)+2\theta_1 E\left(\varepsilon_t\varepsilon_{t-1}\right)+\theta_1^2 E\left(\varepsilon_{t-1}^2\right)\\ &= \left(1+\theta_1^2\right)\sigma_\varepsilon^2.\end{aligned}$$

The autocovariance of a $\mathrm{MA}(1)$ process is obtained from

$$\begin{aligned}\gamma_1 &= E\left[\left(\varepsilon_t+\theta_1\varepsilon_{t-1}\right)x\left(\varepsilon_{t-1}+\theta_1\varepsilon_{t-2}\right)\right]\\ &= E\left[\varepsilon_t\varepsilon_{t-1}+\theta_1\varepsilon_{t-1}\varepsilon_{t-2}+\theta_1\varepsilon_{t-1}\varepsilon_{t-1}+\theta_1^2\varepsilon_{t-1}\varepsilon_{t-2}\right]\\ &= 0+\theta_1.0+\theta_1.\sigma_\varepsilon^2+\theta_1^2.0\end{aligned}$$

as

$$\gamma_k=\begin{cases}\theta_1\sigma_\varepsilon^2\,, & \text{for } k=1\\ 0\quad\;\;, & \text{for } k>1.\end{cases}$$

Since the mean and autocovariances are not functions of time, and $\mathrm{MA}(1)$ process is covariance-stationary regardless of the values of θ_1 (Hamilton, 1994). Autocorrelation of the $\mathrm{MA}(1)$, process can be obtained from γ_k and γ_0 as

$$\rho_k=\begin{cases}\dfrac{\theta_1\sigma_\varepsilon^2}{\left(1+\theta_1^2\right)\sigma_\varepsilon^2}=\dfrac{\theta_1}{\left(1+\theta_1^2\right)} & ,\, k=1\\ 0 & ,\, k>1\end{cases}$$

In Figure 2.7, we simulated n=50 observations from the MA(1) model $Z_t = \varepsilon_t + \theta\varepsilon_{t-1}$, with different theta values respectively 0.7 and -0.5, and $\sigma_\varepsilon^2 = 1$ by using RStudio.

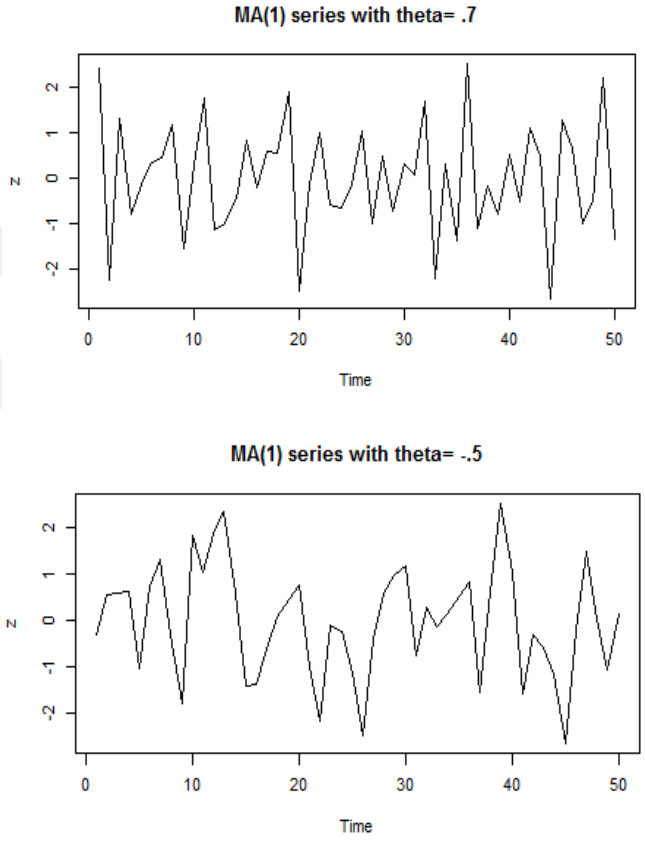


Figure 2.7. Plots of simulated MA(1) series with different theta values

2.5.1.2.(3). A second-order moving average process

A $\mathrm{MA}(2)$ is an expression relating a variable Z_t to current and two previous values of error term and is defined as

$$Z_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$$

where $\varepsilon \sim WN(0,\sigma^2)$, c is a constant term, and θ_1, θ_2 are unknown parameters.

The mean of a $\mathrm{MA}(2)$ process is

$$E\left(Z_t\right)=\mu=c\,.$$

The variance of a $\mathrm{MA}(2)$ process is

$$\begin{aligned}\gamma_0 &= E\left(Z_t-\mu\right)^2\\ &= E\left(\varepsilon_t^2\right)+\theta_1E\left(\varepsilon_t\varepsilon_{t-1}\right)+\theta_2E\left(\varepsilon_t\varepsilon_{t-2}\right)+\theta_1E\left(\varepsilon_t\varepsilon_{t-1}\right)+\theta_1^2E\left(\varepsilon_{t-1}^2\right)\\ &\quad+\theta_1\theta_2E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)+\theta_2E\left(\varepsilon_t\varepsilon_{t-2}\right)+\theta_2\theta_1E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)+\theta_2^2E\left(\varepsilon_{t-2}^2\right)\end{aligned}$$

Since the ε_t's are uncorrelated, the variance can be obtained as

$$\begin{aligned}\gamma_0 &= E\left(\varepsilon_t^2\right)+\theta_1^2E\left(\varepsilon_{t-1}^2\right)+\theta_2^2E\left(\varepsilon_{t-2}^2\right)\\ &= \left(1+\theta_1^2+\theta_2^2\right)\sigma_\varepsilon^2.\end{aligned}$$

The autocovariance of a $\mathrm{MA}(2)$ process is obtained from

$$\begin{aligned}\gamma_1 &= E\left[Z_t.Z_{t-1}\right]\\ &= E\left(\varepsilon_t\varepsilon_{t-1}\right)+\theta_1E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)+\theta_2E\left(\varepsilon_{t-1}\varepsilon_{t-3}\right)+\theta_1E\left(\varepsilon_{t-1}^2\right)+\theta_1^2E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)\\ &\quad+\theta_2\theta_1E\left(\varepsilon_{t-1}\varepsilon_{t-3}\right)+\theta_2E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)+\theta_2\theta_1E\left(\varepsilon_{t-2}^2\right)+\theta_2^2E\left(\varepsilon_{t-2}\varepsilon_{t-3}\right)\\ \gamma_1 &= \theta_1.\sigma_\varepsilon^2+\theta_2\theta_1.\sigma_\varepsilon^2=\left(1+\theta_2\right)\theta_1\sigma_\varepsilon^2.\end{aligned}$$

and

$$\begin{aligned}\gamma_2 &= E\left[Z_t.Z_{t-2}\right]\\ &= \theta_2E\left(\varepsilon_{t-2}^2\right)\\ &= \theta_2\sigma_\varepsilon^2\end{aligned}$$

as

$$\gamma_k=\begin{cases}\left(1+\theta_2\right)\theta_1\sigma_\varepsilon^2 & ,k=1\\ \theta_2\sigma_\varepsilon^2 & ,k=2\\ 0 & ,\mathrm{k}>2.\end{cases}$$

Then, the autocorrelation of the $MA(2)$ process is obtained as

$$\rho_k = \begin{cases} \dfrac{(1+\theta_2)\theta_1\sigma_\varepsilon^2}{(1+\theta_1^2+\theta_2^2)\sigma_\varepsilon^2} = \dfrac{(1+\theta_2)\theta_1}{(1+\theta_1^2+\theta_2^2)}, & k=1 \\ \dfrac{\theta_2\sigma_\varepsilon^2}{(1+\theta_1^2+\theta_2^2)\sigma_\varepsilon^2} = \dfrac{\theta_2}{(1+\theta_1^2+\theta_2^2)}, & k=2 \\ 0 & , k>2 \end{cases}$$

In Figure 2.8, we simulated n=50 observations by using RStudio from the MA(2) model $Z_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$, with theta parameters θ_1 respectively 0.5, 0.6 and θ_2 respectively 0.4,-0.3 and $\sigma_\varepsilon^2 = 1$.

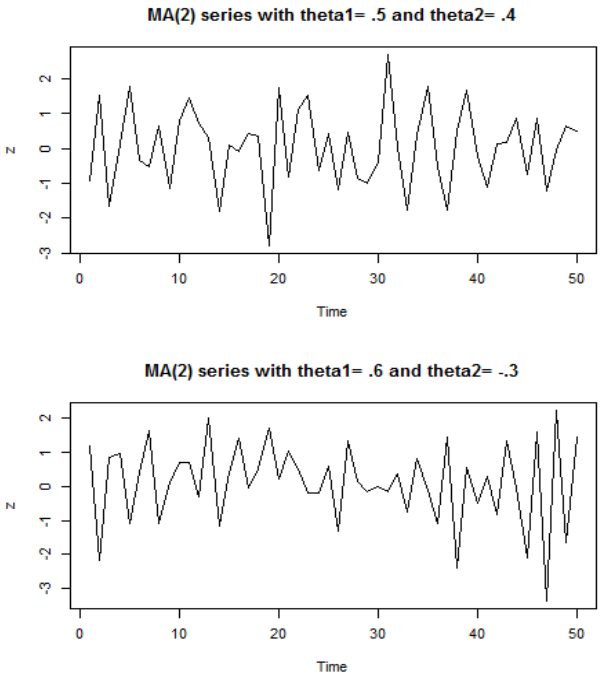


Figure 2.8. Plots of simulated MA (2) series

2.5.1.3. Invertibility

The term invertibility refers to the fact that MA models can be rewritten as an AR model when MA models satisfy invertibility condition.

AR models are *always invertible*, but to be said stationarity models must satisfy some conditions as mentioned in AR process section. Conversely, MA models are *always stationary*, but to be said invertibility models must satisfy some condition also. The condition for invertibility of MA process is counterpart to the condition of stationarity of AR process.

The invertibility condition for a $\mathrm{MA}(q)$ process may be obtained by considering that the model given equation in Eq. (2.20) can be rewritten (assumed no constant, this also means that the constant term has been suppressed by assuming that the variable has been measured as deviations from its mean) by using backshift operator (B) as

$$Z_t = \theta(B)\varepsilon_t \tag{0.21}$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ (Box et al., 1994, p.53). Now rewriting Eq. (2.21) as

$$\varepsilon_t = \theta^{-1}(B) Z_t$$

The invertibility condition for a $\mathrm{MA}(q)$ process is that the roots of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0$$

lie outside the unit circle (Box et al., 1994). This condition is equal to stationarity condition of $\mathrm{AR}(p)$ process as mentioned in the section of *a* p-*th-order autoregressive process.*

Any $\mathrm{MA}(q)$ process can be expressed as an $\mathrm{AR}(\infty)$ process. Suppose that we have a $\mathrm{MA}(1)$ process with no constant,

$$Z_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

or equivalently

$$\varepsilon_t = Z_t - \theta_1 \varepsilon_{t-1}.$$

By substituting ε_t's value step by step, we obtain

$$\varepsilon_t = Z_t - \theta_1 Z_{t-1} + \theta_1^2 Z_{t-2} - \theta_1^3 \varepsilon_{t-3} + \cdots = \sum_{k=0}^{\infty} (-\theta_1)^k Z_{t-k}.$$

Or equivalently we get

$$Z_t = \varepsilon_t + \theta_1 Z_{t-1} - \theta_1^2 Z_{t-2} + \theta_1^3 \varepsilon_{t-3} - \cdots = \varepsilon_t + \sum_{k=1}^{\infty} (-\theta_1)^k Z_{t-k}$$

It turns out that if $|\theta_1| < 1$ then this infinite series converges to a finite value. Then, $\mathrm{MA}(q)$ processes are called invertible. Obviously, Eq. (2.21) is an $\mathrm{AR}(\infty)$ process.

To show an AR process is always invertible without restriction on parameter and can be expressed as an $\mathrm{MA}(\infty)$, now suppose we have $\mathrm{AR}(1)$ process with no constant, then Eq. (2.9) can be rewritten as

$$\begin{aligned} Z_t &= \phi_1(\phi_1 Z_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 Z_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^2(\phi_1 Z_{t-3} + \varepsilon_{t-2}) + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^3 Z_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

sequentially, we obtain

$$\begin{aligned} Z_t &= \phi_1^3(\phi_1 Z_{t-4} + \varepsilon_{t-3}) + \phi_1^2\varepsilon_{t-2} + \phi_1\varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ &= \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} + \phi_1^3\varepsilon_{t-3} + \cdots \end{aligned}$$

which is a $\mathrm{MA}(\infty)$ process.

2.5.1.4. Mixed Autoregressive Moving-Average Process

Together with the autoregressive (AR) model, the moving-average (MA) model is a special case and key component of the more general ARMA and ARIMA models of time series, which have more complicated stochastic structures.

An autoregressive moving average process includes both autoregressive and moving average terms, succinctly ARMA, and is defined as

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \qquad (0.22)$$

or in backshift operator form,

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) Z_t = c + \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)\varepsilon_t.$$

And generally it is convenient to use the more concise form (in terms of deviations from the mean),

$$\phi(B)\bar{Z}_t = \theta(B)\varepsilon_t \qquad (0.23)$$

where $\bar{Z}_t = Z_t - \mu$, $\phi(B)$ and $\theta(B)$ have no common factors and are polynomials of degree p and q in B. We can obviously see that by setting $p \neq 0$ and $q = 0$ we obtain the AR(p) model. Similarly, if we set $p = 0$ and $q \neq 0$ we obtain the MA(q) model.

$\mathrm{ARMA}(p,q)$ process is stationary if the characteristic equation $\phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p=0$ has all roots lying outside the unit circle. $\mathrm{ARMA}(p,q)$ process is invertible if the characteristic equation $\theta(B)=1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q=0$ has all roots lying outside the unit circle. The mean of $\mathrm{ARMA}(p,q)$ process is

$$\begin{aligned}
E(Z_t) &= E\left(c+\phi_1 Z_{t-1}+\phi_2 Z_{t-2}+\cdots+\phi_p Z_{t-p}+\varepsilon+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q}\right)\\
\mu &= c+\phi_1 E(Z_{t-1})+\phi_2 E(Z_{t-2})+\cdots+\phi_p E(Z_{t-p})\\
&\qquad +E(\varepsilon_t)+\theta_1 E(\varepsilon_{t-1})+\cdots+\theta_q E(\varepsilon_{t-q})\\
\mu &= c+\phi_1\mu+\phi_2\mu+\cdots+\phi_p\mu+0+\theta_1.0+\cdots+\theta_q.0\\
\mu &= c/(1-\phi_1-\phi_2-\cdots-\phi_p).
\end{aligned}$$

As it can be seen, the mean of $\mathrm{ARMA}(p,q)$ is the same the one of $\mathrm{AR}(p)$. To find the autocovariance of $\mathrm{ARMA}(p,q)$ process, we apply the relation that the stationary and invertible $\mathrm{ARMA}(p,q)$ process given in Eq. (2.22) has the infinite moving average representation

$$\bar{Z}_t=\psi(B)\varepsilon_t=\sum_{k=0}^{\infty}\psi_k\varepsilon_{t-k}$$

where $\psi(B)=\phi^{-1}(B)\theta(B)$ with the ψ_k weights being absolutely summable. The weights ψ_k are determined from the relation $\phi(B)\psi(B)=\theta(B)$ to satisfy

$$\psi_k=\phi_1\psi_{k-1}+\phi_2\psi_{k-2}+\cdots+\phi_p\psi_{k-p}-\theta_k,\quad k>0$$

with $\psi_0 = 1,\ \psi_k = 0$ for $k < 0,$ and $\theta_k = 0$ for $k > q.$ From this relation the ψ_k weights can readily be computed recursively in terms of the ϕ_i coefficients (Box et al., 1994, p.78).

Now the autocovariance of $\mathrm{ARMA}(p,q)$ process can be obtained by multiplying eq. (2.22) with $\bar{Z}_{t-k}$, then taking expectations give

$$E\left(\bar{Z}_t \bar{Z}_{t-k}\right) = E\left[\left(\phi_1 \bar{Z}_{t-1} + \phi_2 \bar{Z}_{t-2} + \cdots + \phi_p \bar{Z}_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\right) \bar{Z}_{t-k}\right]$$

$$\gamma_k = \phi_1 E\left(\bar{Z}_{t-1} \bar{Z}_{t-k}\right) + \cdots + \phi_p E\left(\bar{Z}_{t-p} \bar{Z}_{t-k}\right) + E\left(\bar{Z}_{t-k} \varepsilon_t\right) + \theta_1 E\left(\bar{Z}_{t-k} \varepsilon_{t-1}\right) + \cdots + \theta_q E\left(\bar{Z}_{t-k} \varepsilon_{t-q}\right)$$

and thus,

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} + \gamma_{z\varepsilon}(k) + \theta_1 \gamma_{z\varepsilon}(k-1) + \cdots + \theta_q \gamma_{z\varepsilon}(k-q) \quad (0.24)$$

where $\gamma_{k-j} = E\left(\bar{Z}_{t-j} \bar{Z}_{t-k}\right)$ and $\gamma_{z\varepsilon}(k) = E\left(\bar{Z}_{t-k} \varepsilon_t\right)$ which is the cross covariance function between Z and ε. Since Z_{t-k} depends only on shocks that have occurred up to time $t-k$ through the infinite moving average representation $\bar{Z}_{t-k} = \psi(B)\varepsilon_{t-k} = \sum_{j=0}^{\infty} \psi_k \varepsilon_{t-k-j}$, it follows that $\gamma_{z\varepsilon}(k) = \psi_{-k}\sigma_\varepsilon^2$ for $k \le 0$, else 0.

Now eq. (2.24) can be rewritten as

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} + \sigma_\varepsilon^2 \left(\theta_k \psi_0 + \theta_{k-1} \psi_1 + \theta_{k-2} \psi_2 + \cdots + \theta_q \psi_{k-q}\right) \quad (0.25)$$

with the convention that $\theta_0 = -1$ (Box et al., 1994).

The variance of $\text{ARMA}(p,q)$ process can be found by setting $k=0$, we have

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma_\varepsilon^2\left(1 + \theta_1\psi_1 + \theta_2\psi_2 + \cdots + \theta_q\psi_q\right)$$

which has to be solved along with the p equations (2.25) for $k=1,2,\ldots,p$ to obtain $\gamma_0,\gamma_1,\ldots,\gamma_p$.

In practice, $\text{ARMA}(1,1)$ model is significantly used. Now we will focus on this model and its structure properties.

2.5.1.4.(1). An ARMA(1,1) process

An $\text{ARMA}(1,1)$ is an expression relating a variable Z_t to one previous value of itself with current and one previous value of error term and also with constant. Model is defined as

$$Z_t = c + \phi_1 Z_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1} \tag{0.26}$$

where $\varepsilon \sim WN(0,\sigma^2)$, c is a constant term, ϕ_1 and θ_1 are unknown parameters.

The mean of an $\text{ARMA}(1,1)$ process is

$$\begin{aligned} E\left(Z_t\right) &= E\left(c + \phi_1 Z_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}\right) \\ \mu &= c + \phi_1 E\left(Z_{t-1}\right) + E\left(\varepsilon_t\right) + \theta_1 E\left(\varepsilon_{t-1}\right) \\ \mu &= c + \phi_1.\mu + 0 + \theta_1.0 \\ \mu &= c/\left(1-\phi_1\right). \end{aligned}$$

The variance of an $\text{ARMA}(1,1)$ process can be found by multiplying Eq. (2.26) with $\overline{Z}_t$ and taking expectations,

$$E\left(\overline{Z}_t\right)^2 = E\left[\overline{Z}_t\left(\phi_1\overline{Z}_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}\right)\right]$$

$$\gamma_0 = \phi_1 E\left(\overline{Z}_t\overline{Z}_{t-1}\right) + E\left(\overline{Z}_t\varepsilon_t\right) + \theta_1 E\left(\overline{Z}_t\varepsilon_{t-1}\right)$$

since $E\left(\overline{Z}_t\overline{Z}_{t-1}\right) = \gamma_1$,

$$E\left(\overline{Z}_t\varepsilon_t\right) = E\left[\left(\phi_1\overline{Z}_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}\right)\varepsilon_t\right] = E\left(\varepsilon_t^2\right) = \sigma_\varepsilon^2$$

and

$$\begin{aligned} E\left(\overline{Z}_t\varepsilon_{t-1}\right) &= E\left[\left(\phi_1\overline{Z}_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}\right)\varepsilon_{t-1}\right] \\ &= \phi_1 E\left(\overline{Z}_{t-1}\varepsilon_{t-1}\right) + \theta_1 E\left(\varepsilon_{t-1}^2\right) \\ &= \phi_1 E\left[\left(\phi_1\overline{Z}_{t-2} + \varepsilon_{t-1} + \theta_1\varepsilon_{t-2}\right)\varepsilon_{t-1}\right] + \theta_1 E\left(\varepsilon_{t-1}^2\right) \\ &= \phi_1 E\left(\varepsilon_{t-1}^2\right) + \theta_1 E\left(\varepsilon_{t-1}^2\right) = \phi_1\sigma_\varepsilon^2 + \theta_1\sigma_\varepsilon^2 \\ &= \left(\phi_1 + \theta_1\right)\sigma_\varepsilon^2 \end{aligned}$$

Now rewriting preceding γ_0, we have

$$\gamma_0 = \phi_1\gamma_1 + \sigma_\varepsilon^2 + \theta_1\left[\left(\phi_1 + \theta_1\right)\sigma_\varepsilon^2\right]$$

$$\gamma_0 = \phi_1\gamma_1 + \sigma_\varepsilon^2\left[1 + \theta_1\left(\phi_1 + \theta_1\right)\right].$$

The autocovariance of an ARMA(1,1) process can be found by multiplying Eq. (2.26) with $\overline{Z}_{t-k}$ and taking expectations,

$$\gamma_1 = \phi_1\gamma_0 + \theta_1\sigma_\varepsilon^2$$

$$\gamma_k = \phi_1\gamma_{k-1} \quad , k \geq 2.$$

Hence, variance and first autocovariance function of the process in terms of parameters is

$$\gamma_0 = \phi_1\left(\phi_1\gamma_0 + \theta_1\sigma_\varepsilon^2\right) + \sigma_\varepsilon^2\left[1 + \theta_1\left(\phi_1 + \theta_1\right)\right]$$

$$\gamma_0 = \phi_1^2\gamma_0 + \phi_1\theta_1\sigma_\varepsilon^2 + \sigma_\varepsilon^2\left[1 + \theta_1\phi_1 + \theta_1^2\right]$$

$$\gamma_0\left(1 - \phi_1^2\right) = \sigma_\varepsilon^2\left[\phi_1\theta_1 + 1 + \theta_1\phi_1 + \theta_1^2\right]$$

$$\gamma_0 = \frac{\sigma_\varepsilon^2\left(1 + 2\theta_1\phi_1 + \theta_1^2\right)}{1 - \phi_1^2}$$

and

$$\gamma_1 = \phi_1\left[\frac{\sigma_\varepsilon^2\left(1+2\theta_1\phi_1+\theta_1^2\right)}{1-\phi_1^2}\right]+\theta_1\sigma_\varepsilon^2$$

$$\gamma_1 = \frac{\sigma_\varepsilon^2\left(\phi_1+2\theta_1\phi_1^2+\phi_1\theta_1^2+\theta_1-\theta_1\phi_1^2\right)}{1-\phi_1^2}$$

$$\gamma_1 = \frac{\sigma_\varepsilon^2\left(\phi_1+\theta_1\phi_1^2+\phi_1\theta_1^2+\theta_1\right)}{1-\phi_1^2}$$

$$\gamma_1 = \frac{\sigma_\varepsilon^2\left(1+\phi_1\theta_1\right)\left(\phi_1+\theta_1\right)}{1-\phi_1^2}$$

Now since we have both variance and autocovariance of ARMA(1,1), autocorrelation of the process can be obtained as

$$\rho_k = \begin{cases}\dfrac{\gamma_k}{\gamma_0} = \dfrac{\left(1+\phi_1\theta_1\right)\left(\phi_1+\theta_1\right)}{1+2\theta_1\phi_1+\theta_1^2}\phi_1^{k-1} & , k \geq 1\end{cases}$$

In Figure 2.9, we shows simulated n=50 observations by using RStudio from the ARMA(1,1) model with phi parameter $\phi_1 = 0.5$ and theta parameter $\theta_1 = 0.6$, and $\sigma_\varepsilon^2 = 1$.

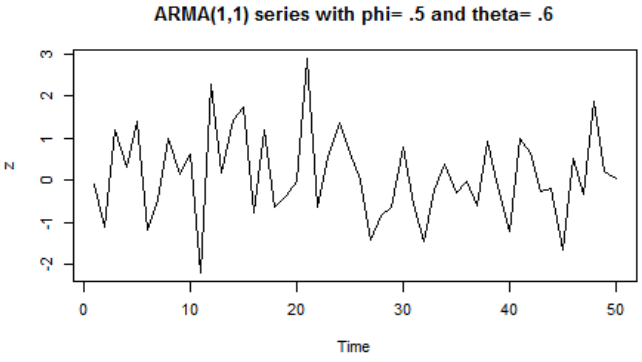


Figure 2.9. Plot of simulated ARMA (1,1) series

2.5.2. Non-Stationary Time Series Models

In real life, many time series does not have a constant mean over time and mostly have trend which means series is nonstationary. Models that show such behaves can be made to be stationary by some suitable difference of the process as well as the logarithm or power transformation which are also useful methods in certain circumstances but nonlinear transformation. Now we focus on the properties of this type of model which the *d*-th difference is a stationary mixed autoregressive moving average process and is called autoregressive integrated moving average, succinctly ARIMA, process.

2.5.2.1. Autoregressive Integrated Moving Average Process

ARIMA process includes both autoregressive and moving average terms together with difference operator (degree of 1 or at most 2), and $ARIMA(p,d,q)$ model is defined as,

$$\nabla^d Z_t = c + \phi_1 \nabla^d Z_{t-1} + \phi_2 \nabla^d Z_{t-2} + \cdots + \phi_p \nabla^d Z_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{0.27}$$

where ∇ is difference operator and d is the order of difference, and others are same as given previously. In backshift operator form,

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) \nabla^d Z_t = c + \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) \varepsilon_t$$

and suppose $\nabla^d Z_t = Y_t$ which is differenced series, it is convenient to use the more concise form,

$$\phi(B) Y_t = c + \theta(B) \varepsilon_t.$$

Stationarity and invertibility conditions of the model are the same as in ARMA models.

2.6. Model Selection

There are several methods to select which model is more appropriate for a given time series. To learn how the identification of the model such as specifying ARMA order by means of the plot of ACF and PACF see Box et al 1994, Pena et al 2001.

One method for model selection, which is used in application section of this thesis, is to use information criteria such as AIC proposed by Akaike or BIC (also known SIC) of Schwarz. These two criteria are given by

$$AIC_{p,q} = \ln\left(\hat{\sigma}_{\varepsilon}^{2}\right) + r\frac{2}{n}$$

and

$$BIC_{p,q} = \ln\left(\hat{\sigma}_{\varepsilon}^{2}\right) + r\frac{\ln(n)}{n}$$

where $\hat{\sigma}_{\varepsilon}^{2}$ is the maximum likelihood estimate of $\hat{\sigma}_{\varepsilon}^{2}$ and $r = p + q$ is the number of parameters that are estimated in the model, and n is the length of the series. Among finite set of models, the model with the lowest AIC or BIC value is preferred.

3. OUTLIERS in TIME SERIES

Fox (1972) have done the first study about outliers in the time series. He defined possible two types of outliers which can occur in the autoregressive (AR) time series as additive outlier (AO) and innovation outlier (IO). Chang (1982) and Chang and Tiao (1983) extended Fox's test to ARIMA models and proposed an iterative procedure which detects multiple outliers (Tsay, 1988). After two outlier types presented by Fox (1972), Tsay (1988) introduced new outlier types as level shifts (LS), temporary change (TC). He also defined variance change (VC) as outlier type, which is mostly regarded as structure change and is not focused in this study.

Guttman and Tiao (1978), Miller (1980), and Chang (1982) showed that the effect of such interventions (outliers) may cause serious bias in estimating autocorrelations, partial autocorrelations, and autoregressive moving average (ARMA) parameters. It is, therefore, important to be able to identify these interventions and delete their effects from the data set to better understand the underlying structure of the series. If location of intervention point T is known, the effect of it can frequently be explained by using the intervention analysis techniques which are proposed by Box and Tiao (1975). In practice, however, information about the timing is seldom available (Chang et al., 1998).

In the literature, there exist many time series models and different outlier types according to models. In this study, outliers in ARIMA models are considered.

3.1. Additive Outlier (AO)

Additive outlier (or observational outlier) is regarded as a single observation that occurs at time $t=T$. It is a surprisingly large or small value and does not affect subsequent observations. Typical examples of AO are the recording error, crisis or strike.

AO model is defined as

$$Y_t = Z_t + \omega_A I_t^{(T)} \tag{0.28}$$

where Y_t is the observed time series and Z_t is the unobservable original series which is free of the impact of outliers and follows ARMA process as given in Eq. (2.22), and ω_A represents the magnitude of the additive outlier, and $I_t^{(T)}$ is an indicator (or impulse) variable which signifies the occurrence of the outlier such that

$$I_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}.$$

We simulated n=100 observations by using RStudio, which is shown in Figure 3.1., from the AR(1) model with an AO of size 6 at t=75, phi parameter $\phi_1 = -0.5$ and $\sigma_\varepsilon^2 = 1$. Red point in the plot indicates the occurrence of AO time.

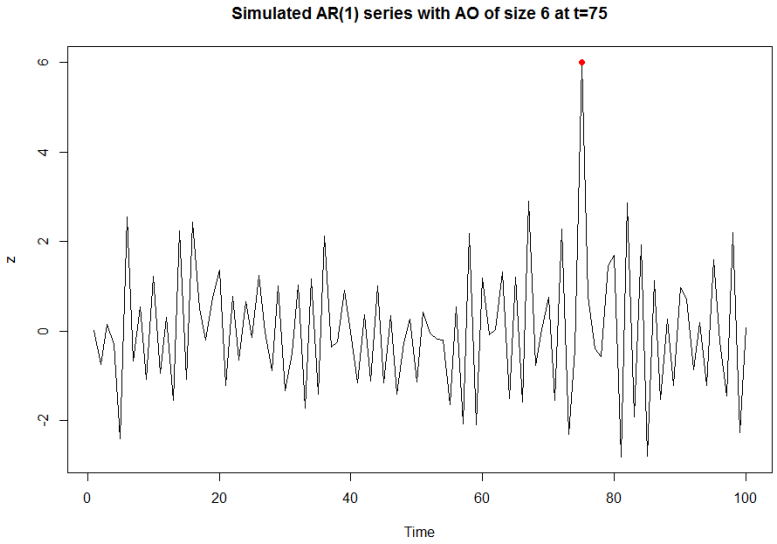


Figure 3.1. Plot of simulated AR(1) with an AO at t=75

3.2. Innovational Outlier (IO)

Innovational outlier affects the subsequent observations starting from its position. Although there is no general acceptance about the number of observations in IO, we usually encounter two-three observations.

IO model is defined as

$$Y_t = Z_t + \frac{\theta(B)}{\phi(B)}\omega_I I_t^{(T)} \tag{3.2}$$

where Y_t and Z_t are as defined in Section 3.1., and ω_I represents the magnitude of the innovational outlier, and $I_t^{(T)}$ is an indicator (or impulse) variable which signifies the occurrence of the outlier such that

$$I_t^{(T)} = \begin{cases} 1, & t \geq T \\ 0, & t < T \end{cases}.$$

Thus the AO model can be regarded as a gross error model, because only the level of the T*th* observation is affected. On the other hand, the IO model represents a disturbance on the white noise series ε_t at time point *T*, which influences the process Z_t on $Z_t, Z_{t+1}, \ldots$, through the dynamic structure $\theta(B)/\phi(B)$ (Tsay, 1986).

We simulated n=100 observations by using RStudio, which is shown in Figure 3.2., from the AR(1) model with an IO of size 5 at t=75, phi parameter $\phi_1 = -0.5$ and $\sigma_\varepsilon^2 = 1$. Red point in the plot indicates the occurrence of innovational outlier time and black points are subsequent observations which are effected from the outlier.

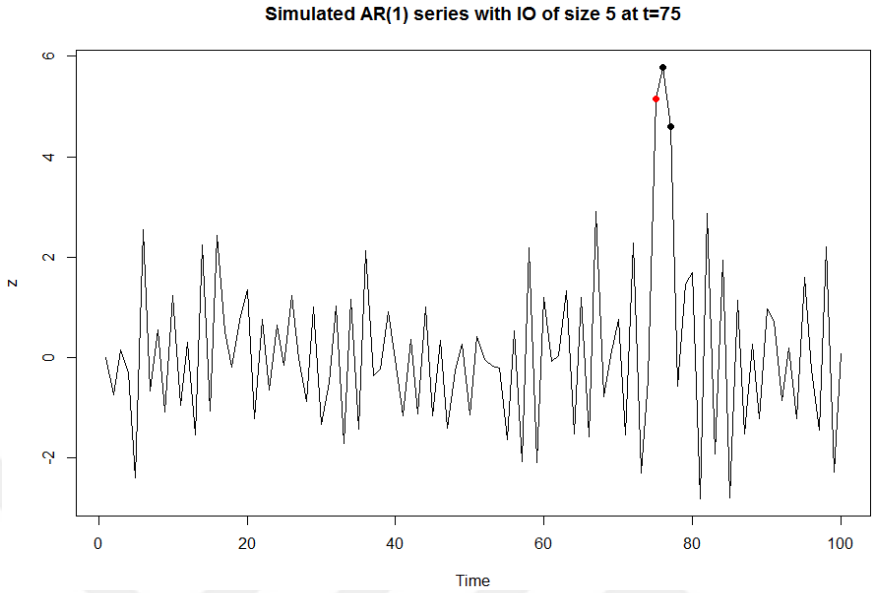


Figure 3.2. Plot of simulated AR(1) with an IO at t=75

3.3. Level Chance (LC)

Level change (or level shift) is regarded as a change in the mean of series, and starts a certain point of series.

LC model is defined as

$$Y_t = Z_t + \frac{1}{1-B}\omega_L I_t^{(T)} \tag{3.3}$$

where Y_t, Z_t and $I_t^{(T)}$ are as defined in Section 3.1., and ω_L represents the magnitude of the level change.

The model says that a level shift of magnitude ω_L occurs at time $t=T$ and the change is permanent (Tsay, 1988).

We simulated n=100 observations by using RStudio, which is shown in Figure 3.3., from the AR(1) model with an LC of size 5 at t=75, phi parameter

$\phi_1 = -0.5$ and $\sigma_\varepsilon^2 = 1$. Red point in the plot indicates the occurrence of level change and black points are consecutive observations effected from the outlier.

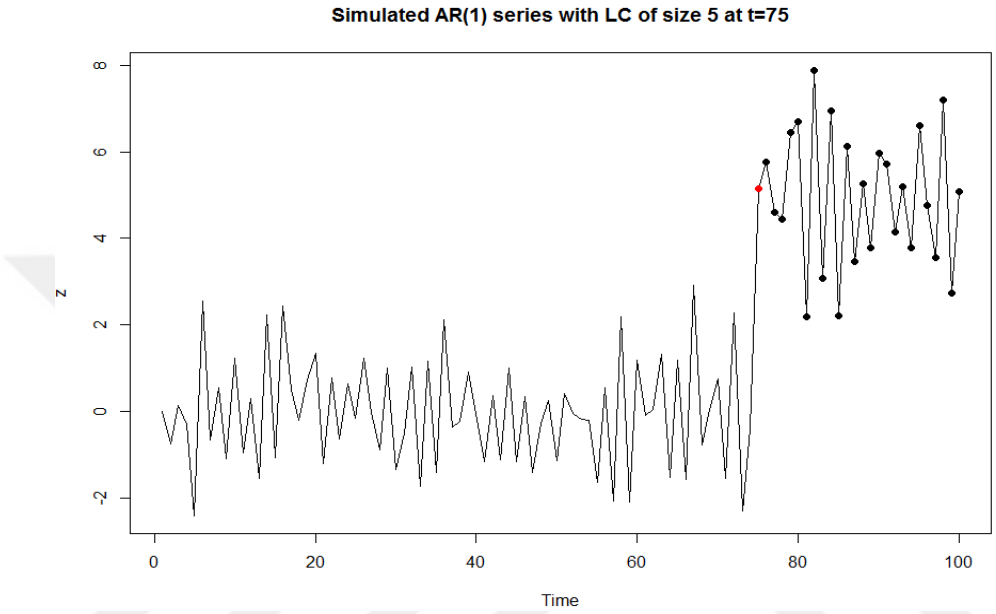


Figure 3.3. Plot of simulated AR(1) with an LC at t=75

3.4. Transient Chance (TC)

Temporary change affects the subsequent observations starting from its position.

TC model is defined as

$$Y_t = Z_t + \frac{1}{1-\delta B}\omega_{TC} I_t^{(T)}, \quad 0 < \delta < 1 \tag{3.4}$$

where Y_t, Z_t and $I_t^{(T)}$ are as defined in Section 3.1., and ω_{TC} represents the magnitude of the temporary change. The AO and LS are two boundary cases of TC, where $\delta = 0$ and $\delta = 1$.

The model describes a disturbance that affects Z_t for all $t \geq T$. However, the effect decays exponentially with rate δ and an initial impact ω_{TC}. Since $0 < \delta < 1$, the effect disappears eventually. For this reason, this model is referred to as a transient change (TC) model (Tsay, 1988).

For a TC, the outlier produces an initial effect ω_{TC} at time *T*, and this effect dies out gradually with time. The parameter δ is designed to model the pace of the dynamic dampening effect. In practice, the value of δ can be specified by the analyst. We recommend that $\delta = .7$ be used to identify a TC (Chen and Liu, 1993).

We simulated n=100 observations by using RStudio, which is shown in Figure 3.4., from the AR(1) model with an LC of size 5 at t=70, phi parameter $\phi_1 = -0.5$ and $\sigma_\varepsilon^2 = 1$. Red point in the plot indicates the occurrence of temporary change and black points are consecutive observations effected from the outlier.

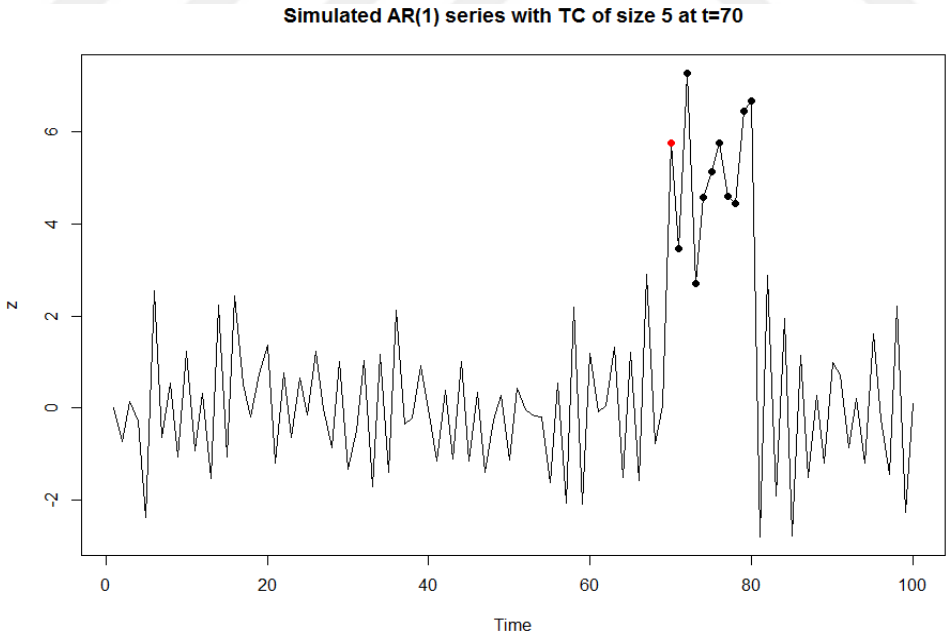


Figure 3.4. Plot of simulated AR(1) with an TC at t=70

It is seen from the plot that there is a change in the mean of the series at temporary change interval and this change is not permanent. The mean of the series comes back regular level after the TC interval.

3.5. The General Outlier Model

Considering four types of outlier model, the general outlier model can be written as

$$Y_t = Z_t + V(B)\omega_i I_t^{(T)} \tag{3.5}$$

where Y_t, Z_t and $I_t^{(T)}$ are as defined in Section 3.1., and ω_i represents the magnitude of outlier (outlier size) and subscript denotes the outlier type $i=A$ for AO, $i=I$ for IO, $i=L$ for LC and $i=TC$ for TC, and $V(B)$ defines the outlier type such that

$$V(B) = \begin{cases} 1 & \text{, for AO} \\ \dfrac{1}{\pi(B)} & \text{, for IO} \\ \dfrac{1}{1-B} & \text{, for LC} \\ \dfrac{1}{1-\delta B} & \text{, for TC} \end{cases}$$

where the value of δ is prespecified such that $0 < \delta < 1$ and $\pi(B) = \phi(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$.

Now let $e_t = \pi(B)Y_t$, then Eq. (3.5) can be rewritten in terms of ε_t as

$$e_t = \omega_i x_t + \varepsilon_t \tag{3.6}$$

where ω_i is as defined in Eq. (3.5) and x_t is defined such that

$$x_t = \begin{cases} \pi(B) I_t^{(T)} & \text{, for AO} \\ I_t^{(T)} & \text{, for IO} \\ \dfrac{\pi(B) I_t^{(T)}}{1-B} & \text{, for LC} \\ \dfrac{\pi(B) I_t^{(T)}}{1-\delta B} & \text{, for TC.} \end{cases}$$

The information about an IO is contained in the residual ε_t at that particular point *T*, whereas that of an AO is scattered over string of residuals $\varepsilon_t, \varepsilon_{t+1}, \ldots$ (Chang et al., 1988).

3.6. Estimation of Outlier Effects

To examine the effects of outliers on the estimated residuals, it is assumed that the time series parameters are known. The least squares estimate of the magnitude ω of a single outlier at $t=T$ can be estimated from Eq. (3.6) as

$$\hat{\omega}_i = \frac{\sum_{j=1}^{n} e_t x_{t,j}}{\sum_{j=1}^{n} x_{t,j}^2}$$

and the variance of the magnitude of outlier is

$$Var\left(\hat{\omega}_i\right) = \sigma_\varepsilon^2 \left(\sum\nolimits_{j=1}^{n} x_{t,j}^2\right)^{-1} .$$

In the AO case, the estimation leads to

$$\hat{\omega}_A = \rho_A^2 \pi\left(F\right) e_T \tag{3.7}$$

where $\pi\left(F\right) = \left(1 - \pi_1 F - \pi_2 F^2 - \cdots - \pi_{n-T} F^{n-T}\right)$ and F is the forward operator defined in Section 2.2., and $\rho_A^2 = \left(1 + \pi_1^2 + \cdots + \pi_{n-T}^2\right)^{-1}$. The variance is $Var\left(\hat{\omega}_A\right) = \rho_A^2 \sigma_\varepsilon^2$.

In the IO case, the only residual that has information about the outlier size is the one at *T*, and so, as one would expect, the estimate is

$$\hat{\omega}_I = e_T \tag{3.8}$$

and the variance is $Var\left(\hat{\omega}_I\right) = \sigma_\varepsilon^2$.

In the LS case, all the residuals after the shift have information, and we would expect that the estimate combines all of them in some optimal way. The estimate is

$$\hat{\omega}_L = \rho_L^2 l\left(F\right) e_T \tag{3.9}$$

where $l\left(B\right) = \pi\left(B\right) / \left(1 - B\right)$ and $\rho_L^2 = \left(1 + l_1^2 + \cdots + l_{n-T}^2\right)^{-1}$, where l_i are the coefficients of $l\left(B\right)$. The variance is $Var\left(\hat{\omega}_L\right) = \rho_L^2 \sigma_\varepsilon^2$.

Finally, in the TC case

$$\hat{\omega}_{TC} = \rho_{TC}^2 \beta(F) e_T \tag{3.10}$$

where $\rho_{TC}^2 = \left(1 + \beta_1^2 + \cdots + \beta_{n-T}^2\right)^{-1}$, in which β_i are the coefficients of $\beta(B) = \pi(B)/(1-\delta B)$. The variance is $Var(\hat{\omega}_{TC}) = \rho_{TC}^2 \sigma_\varepsilon^2$ (Pena et al. 2001).

3.7. Testing for an Outlier

To test whether or not one outlier of known type has occurred at specific time *T*, the standard test is

$$H_0 : \omega_j = 0$$
$$H_j : \omega_j \neq 0$$

where $j = I, A, L, TC$ according to outlier type that to be tested. If we know the parameters, then e_t, which is given by (3.6), can be used and the test equals to test the slope of a simple regression model. As we quite know, the likelihood ratio criterion cause comparing the estimated parameter to its standard error, and thus the statistic of test is

$$\lambda_{j,t} = \frac{\hat{\omega}_{j,t}}{\rho_{j,t}\sigma_\varepsilon} \tag{3.11}$$

where $\rho_{I,T} = 1$ and $\lambda_{j,t}$ has student-t distribution. If occurrence time of the outlier is not known, Fox (1972) offered using the likelihood ratio test $\eta_t = \max_{1 \le t \le n}\left(\lambda_{j,t}\right)$. However, because of the correlation between $\lambda_{j,t}$ the sample distribution of η_t is complex. The percentiles of the distribution can be found by

simulation, and they were acquired by Chang et al. (1988). Ljung (1993) have suggested some approximations to this distribution (Pena et al. 2001).

The decision rule for the test is that if $\max_{1\le t\le n}\left|\lambda_{j,t}\right|\ge C$ then an outlier is detected at time t and type $j=I,A,L,T$, where C is a predetermined constant and takes values 3, 3.5 and 4 according to the sensitivity level. For high sensitivity (%99 confidence level) $C=3$, for middle sensitivity (%95 confidence level) $C=3,5$ and for low sensitivity (%90 confidence level) $C=4$.

In practice, the time series parameters are unknown and estimated from the data. The unknown parameters of the test statistics, such as $\rho_{j,t}$ and σ_ε, are replaced by some consistent estimates. Under the null hypothesis that there is no outlier and the assumption that T is known, the test statistics of (3.11) are asymptotically distributed as N(0, 1) (Tsay, 1988).

3.7.1. Confusion between IO and LC

Sanchez and Pena (2003) mentioned about the confusion between IO and LC when the series has a level shift. Balke (1993) showed that in the present of a level shift in a stationary time series, LC is usually identified as an IO.

To understand the confusion between IO and LC, comparing the test statistics for these two outliers can be helpful.

The statistics for testing for an IO is

$$\hat{\lambda}_{I,T}=\frac{\hat{e}_T}{\hat{\sigma}_\varepsilon}$$

whereas for an LS is

$$\hat{\lambda}_{\mathrm{L},T} = \frac{\hat{e}_T - \sum_{i=1}^{n-T} \hat{l}_i \hat{e}_{T+i}}{\hat{\sigma}_\varepsilon \sqrt{\left(1 + \sum_{i=1}^{n-T} \hat{l}_i^2\right)}}$$

where $\hat{l}_1 = -1 + \hat{\pi}_1, \hat{l}_2 = -1 + \hat{\pi}_1 + \hat{\pi}_2, \ldots, \hat{l}_{n-T} = -1 + \sum_{i=1}^{n-T} \hat{\pi}_i$. Then the relationship between $\hat{\lambda}_{I,T}$ and $\hat{\lambda}_{L,T}$ is:

$$\hat{\lambda}_{\mathrm{L},T} = \frac{\hat{\lambda}_{\mathrm{I},T} + \frac{\sum_{i=1}^{n-T}\left(\hat{e}_{T+i}\left(1 - \sum_{j=1}^{i} \hat{\pi}_j\right)\right)}{\hat{\sigma}_\varepsilon}}{\sqrt{\left(1 + \sum_{i=1}^{n-T}\left(1 - \sum_{j=1}^{i} \hat{\pi}_j\right)^2\right)}}$$

For an invertible ARMA model, under H_0 (no outliers), when $t = T$ is not close to the end of the series, for large n, the second term will go to zero and the likelihood ratio for level shifts, $\hat{\lambda}_{L,T}$, is expected to be smaller than the likelihood ratio for innovational outliers, $\hat{\lambda}_{I,T}$. This result suggests that the critical values under the null hypothesis of these two statistics can be quite different (Pena et al. 2001).

In order to check the behaviour of $\hat{\lambda}_{\mathrm{L},T}$ and $\hat{\lambda}_{I,T}$ in finite samples, Sanchez and Pena (2003) have carried out a large simulation study and they recommended the critical values for $\hat{\lambda}_{\mathrm{L},T}$ and $\hat{\lambda}_{I,T}$ shown in Table 3.1.

Table 3.1. Recommended critical values for the outlier type of IO and LS by Sanchez and Pena (2003)

n	IO	LS-ST	LS-NST
50	3.10	2.60	3.35
100	3.35	2.75	3.55
250	3.65	2.90	3.75

ST = stationary models, NST = nonstationary models.

As can be noticed from Table 3.1, the critical values for IO are slightly greater than LC for stationary models, whereas critical values for LC are greater than IO for nonstationary models.

3.8. Iterative Procedure for Outlier Detection

While testing for outliers in time series, more than one type of outlier can be detected at time *t*. In order to cope with this problem and to detect the time at which outlier occurs, and also the type of outlier at that time, an iterative procedure for outlier detection is proposed by Chang and Tiao (1983) and Chang et al. (1988) which is only concerned with AO and IO. Then the procedure was generalized by Tsay (1988) to detect LS and TC. Chen and Liu (1993) made some improvement on this procedure which is widely used so far. Sanchez and Pena (2003) also made improvement on the procedure which keeps the powerful features of previous methods but improves the initial parameter estimate, avoids the confusion between innovative outliers and level shifts. A combine of this iterative procedure for outlier detection runs as follows.

Stage I (Initial Parameter Estimation):

1. In the first stage, a model is fitted by maximum likelihood (ML) to the time series assuming that there is no outlier.

2. From the estimated model, obtain the residuals ε_t.
3. Compute the median absolute deviation (MAD) for variance estimate of residuals σ_ε, from

$$\hat{\sigma}_\varepsilon = 1.483 \times median\left\{\left|\hat{\varepsilon}_t - median\left(\hat{\varepsilon}\right)\right|\right\}.$$

Stage II (Outlier Detection and Removing Effects):

4. In the second stage, for $t = 1, 2, \ldots, n$, compute the likelihood ratios $\hat{\lambda}_{\mathrm{A},t}, \hat{\lambda}_{\mathrm{I},t}, \hat{\lambda}_{\mathrm{O},t}$ and $\hat{\lambda}_{\mathrm{TC},t}$ from Eq. (3.11). In other words, calculate all the tau-statistics for each type of outlier and for every time point.
5. For all $t = 1, 2, \ldots, n$, the statistic $\eta_t = \max\left\{\left|\hat{\lambda}_{\mathrm{A},t}\right|, \left|\hat{\lambda}_{I,t}\right|, \left|\hat{\lambda}_{\mathrm{L},t}\right|, \left|\hat{\lambda}_{TC,t}\right|\right\}$ is kept. In other words, for each time point, select the one that has maximum tau statistic in absolute value.
6. All η_t that are greater than C in absolute value are selected and defined as outlier. Here, C is predetermined critical value and usually taken 3 for %90 confidence level(CL), 3.5 for %95 CL or 4 for %99 CL. In this step, there are some rules ;
   i. If same type of outliers detected at consecutive time points, select the one that has maximum tau statistic in absolute value.
   ii. If there are both LS and another type of outlier at consecutive time points, select the one that has maximum tau statistic in absolute value.
7. If there is no outlier detected, then stop and the result is that this time series does not have any outliers. When an outlier is detected, delete the effect of the outlier from the residuals according to the type of outlier, such

$$\breve{\varepsilon}_t = \begin{cases} \hat{\varepsilon}_t - \hat{\omega}_I & , \ t = T \ \text{ for IO} \\ \hat{\varepsilon}_t - \hat{\omega}_A \hat{\pi}(B) I_t^{(T)} & , \ t = T \ \text{ for AO} \\ \hat{\varepsilon}_t - \hat{\omega}_L \hat{l}(B) I_t^{(T)} & , \ t \geq T \ \text{ for LC} \\ \hat{\varepsilon}_t - \hat{\omega}_{TC} \hat{\beta}(B) I_t^{(T)} & , \ t \geq T \ \text{ for TC.} \end{cases}$$

Stage III (Final Parameter Estimates):

8. Using the new residuals $\breve{\varepsilon}_t$ of adjusted series, which is obtained in step 7, estimate parameters of the final model.
9. Steps 1 through 8 are repeated until no more outliers are found in final time series.

3.9. Influential Observations

An observation may not be identified as outlier, but it can still effect the parameter estimations of the model. This kind of observation is called as influential observation. It is important to detect, if any, these observations in time series.

Suggested statistics to measure the effect of an observation are usually adapted from that are used in regression analysis.

Bruce and Martin (1989) proposed diagnostics for ARIMA model fitting for time series formed by deleting observations from the data and measuring the change in the estimates of the parameters.

Abraham and Chuang (1989) have presented tests, which are based on the deletion of the observation and similar with the one that Draper and John (1981) suggested for regression situation, to detect influential observations and outliers in time series.

Pena (1990) studied the influential observations on the coefficients of ARIMA model and introduced influence statistics based on Mahalanobis distance. Suggested measure of influence on the parameter values of an ARIMA model is

$$D_2(\mathrm{T}) = \frac{\left(\hat{\pi} - \hat{\pi}_T\right)^{'}\left(X_Z^{'} X_Z\right)\left(\hat{\pi} - \hat{\pi}_T\right)}{h\hat{\sigma}_\varepsilon^2}$$

where $\pi(B) = \phi(B)(1-B)^d \theta(B)^{-1}$, $\hat{\pi}$ is the MLE of the parameter vector π assuming no outlier, and $\hat{\pi}_T$ is the MLE of π_T assuming that observation at t=T is missing, and $\hat{\sigma}_\varepsilon^2$ is the variance-covariance matrix of the $\hat{\pi}$ estimated vector, and h is the number of parameters, and X_Z to the matrix

$$X_Z = \begin{pmatrix} Z_h & Z_{h-1} & \cdots & Z_1 \\ Z_{h+1} & Z_h & \cdots & Z_2 \\ \vdots & \vdots & \ddots & \vdots \\ Z_{t-1} & Z_{t-2} & \cdots & Z_{t-h} \end{pmatrix}.$$

After computing $D_2(T)$ statistic for all of the individual points, Pena suggested selecting only the $\alpha\%$ points (between 10% and 25%) with highest influence.

Hui and Lee (1992) introduced a leave-one-out procedure, which has been widely used in regression analysis, to identify influential points. They proposed three measures to quantify the influence of z_T on the estimates as follows.

i. Parzen mean square percentage error statistic:

ii.

$$d_1(\mathrm{T}) = \frac{\sum_{j=0}^{t/2}\left[\hat{I}_T(\omega_j) - \hat{I}(\omega_j)\right]^2}{\hat{I}(\omega_j)^2}$$

where $\hat{I}_T(\omega)$ is obtained from the equation

$$f_y(\omega_j) = \frac{1}{2\pi n}\sum_{k=0}^{n-1}\left[f_z\left(\omega_j - \omega_k\right) I_g\left(\omega_k\right)\right]$$

where

$$I_g(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} g_t \exp\left(\text{it } \omega\right)\right|^2$$

after deleting z_T from data set, and $\hat{I}\left(\omega_j\right)$ is computed over the discrete set of frequencies $\omega_j = 2\pi j/n,\ j = 0, \pm 1, \pm 2, \ldots, n/2.$

iii. Log transformation (variance stabilizing) statistic:

iv.

$$d_2(\mathrm{T}) = \frac{\sum_{j=0}^{t/2}\left[\log \hat{I}_T\left(\omega_j\right) - \log \hat{I}\left(\omega_j\right)\right]^2}{\log \hat{I}\left(\omega_j\right)^2}$$

v. Grenander and Rosenblatt statistic:

$$d_3(\mathrm{T}) = \max\left[\frac{1}{\sqrt{n}}\left|\hat{F}_T\left(\omega_j\right) - \hat{F}\left(\omega_j\right)\right|\right]$$

where $\hat{F}\left(\omega_j\right) = \sum_{k=0}^{j}\hat{I}\left(\omega_k\right).$

To assess the influential observation, they pointed out that a substantive large d_i value indicates a significant influence on the estimated spectrum and also diagnostic plots of these statistics on deleting one observation at a time will help us to identify suspected observations.

Lin and Lee (2001) adapted conditional Cook's distance, which is suggested by Lawrance and used in regression analysis, to the autoregressive models. To understand how they adapted, consider AR(p) model;

$$Z = X\phi + \varepsilon$$

where $Z^T = \left(z_{p+1},\ldots,z_n\right)$, $\phi^T = \left(\phi_1,\ldots,\phi_p\right)$, $\varepsilon^T = \left(\varepsilon_{p+1},\ldots,\varepsilon_n\right)$, and

$$X = \begin{pmatrix} z_p & z_{p-1} & \cdots & z_1 \\ z_{p+1} & z_p & \cdots & z_2 \\ \vdots & \vdots & \ddots & \vdots \\ z_{n-1} & z_{n-2} & \cdots & z_{n-p} \end{pmatrix}$$

as it is well known, the conditional least square estimate of ϕ is given by

$$\hat{\phi} = \left(X^T X\right)^{-1} X^T Z$$

and the fitted values are

$$\hat{Z} = X\hat{\phi} = HZ$$

where $H = X\left(X^T X\right)^{-1} X^T$ and the residuals are $e = \left(1-H\right)Z$. Cook's distance corresponding single deletion is

$$C_i = \left\{\hat{\phi}_{(i)} - \hat{\phi}\right\}^T \left(X^T X\right)\left\{\hat{\phi}_{(i)} - \hat{\phi}\right\} \Big/ ps^2$$

where $s^2 = \sum_{i=1}^{n} e_i^2 \big/ \left(n-p\right)$, and subscript (i) denotes the deletion of the i-th observation. Conditional Cook's distance to be the Cook's distance for the i-th observation after deletion of the j-th observation:

$$C_{i(j)} = \left\{\hat{\phi}_{(i,j)} - \hat{\phi}_{(j)}\right\}^T \left(X_{(j)}^T X_{(j)}\right)\left\{\hat{\phi}_{(i,j)} - \hat{\phi}_{(j)}\right\} \Big/ ps^2$$

which can be written as

$$C_{i(j)} = \frac{\left\{h_i / \left(1-h_i\right) + \tilde{h}_{ij}^2\right\} r_i^2 \left\{1 + \tilde{h}_{ij}\left(r_i / r_j\right)\right\}^2}{\left(1-\tilde{h}_{ij}^2\right)^2 p}$$

where $\tilde{h}_{ij} = h_{ij} \Big/ \sqrt{\left(1-h_i\right)\left(1-h_j\right)}$ denotes the standardized joint leverage, $h_{ij} = x_i \left(X^T X\right)^{-1} x_j$ is the (i,j)th off-diagonal element of H, and $r_i = e \Big/ \sqrt{s\left(1-h_i\right)}$ defines the i-th standardized residual.

To assess the influential observation in AR(p) models, Lin and Lee suggested selecting the $\max C_{i(j)}$, then compare with appropriate critical values. In this manner, the exact sampling distribution of this statistic is unknown. That's why they appealed to large-sample theory. Consequently, they came up with that for a given j, $\max C_{i(j)}$ has an asymptotic $D_j \chi_1^2$ distribution with the constant D_j which is defined as

$$D_j = \frac{\left[h_j \Big/ \left(1-h_j\right) + h_{jj(j)}\right] + \left(1 + h_{jj(j)}\right)}{p\left(1-h_{jj(j)}\right)^2 \left(1-h_j\right)}.$$

They pointed out that the χ^2 limits should not be interpreted as significance tests but rather like warning limits indicating the particular observation should be further examined closely.

They also suggested $M_{i(j)} = C_{i(j)} \big/ C_i$ ratio, which provides a convenient measure of the masking or boosting effect of case at time *i* by case at time *j*. When $M_{i(j)} > 1$, it indicates the masking of observation i by observation j, and boosting when $0 \le M_{i(j)} \le 1$.

4. APPLICATION

In this section, both a simulated time series and a real time series are considered to illustrate the outlier detection procedure. The procedure is carried out step by step and interprets about the related figures and tables are presented.

4.1. Outlier Detection with a Simulated Series

We generated 150 observations from ARIMA(1,0,1) process with phi parameter $\phi_1 = 0.5$ and theta parameter $\theta_1 = -0.4$, and $\sigma_\varepsilon^2 = 1$ by using R-Studio. Simulated time series is given in Appendix-1 and its graph is shown in Figure 4.1.

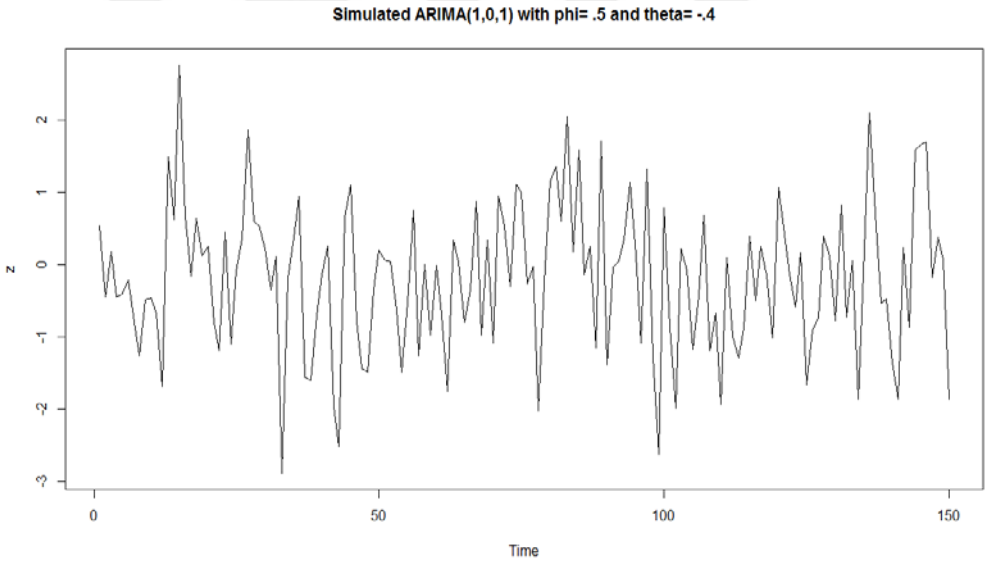


Figure 4.1. Plot of simulated ARIMA (1,0,1) with phi=.5 and theta=-.4

As can be seen from Figure 4.1., there seems no abnormality in the simulated time series. It is just suspected from the observations 15, 33 and 99 and ADF test is performed to test stationarity, and according to results, this simulated time series is stationary. Outlier detection procedure is also performed, and no

outlier is detected. Then, we estimated model parameters for this simulated time series which is given in Table 4.1.

Table 4.1. Estimated parameters for the simulated time series

	AR(1)	**MA(1)**
Coefficients	0.4673	-0.3222
Standard Error	0.4638	0.4991
Sigma^2 estimated as 1.0602: log likelihood=-216.38		
AIC=438.75	**AICc**=438.92	**BIC**=447.79

Then, in order to illustrate how to detect outlier in time series, we introduced three outliers to the simulated time series. We added an AO at $t=15$ of size 4, and at $t=33$ an AO of size -3, and at $t=110$ a LS of size 5. Figure 4.2 shows the plot of this contaminated time series.

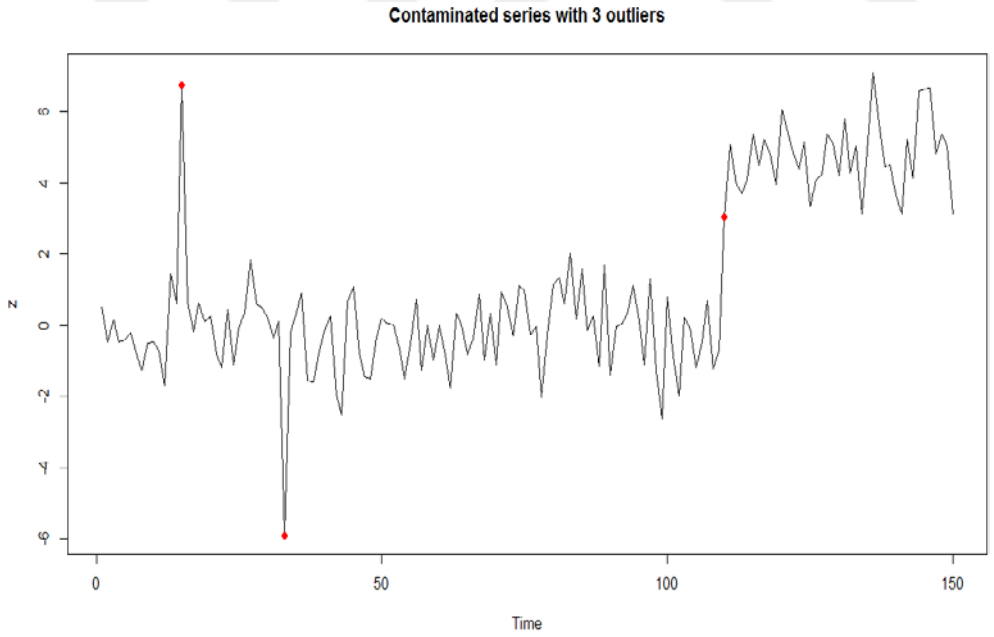


Figure 4.2. Contaminated time series with AO at t=15,33 and a LC at t=110

As can be seen from Figure 4.2., contaminated time series has two outliers and a permanent level shift. To test stationarity, we have performed ADF test and result is given in Table 4.3.

Table 4.2. ADF test result for contaminated time series

Test Statistic	p-value
-2.3654	0.4242

We tested null hypothesis that contaminated time series is not stationary against alternative hypothesis that contaminated time series is stationary, and according to test results, null hypothesis failed to reject with %95 confidence level since $p = 0.4242 > \alpha = 0.05$. As it is noticed, the presence of outliers in time series affects the series' stationarity.

Outlier detection procedure, which is given in section 3.8, is implemented to the contaminated time series step by step, and steps of the procedure are as follow:

Step 1: ARIMA(0,1,1) model is fitted to the contaminated time series by using auto.arima function of RStudio and result is given in Table 4.3.

Table 4.3. Fitted ARIMA to the contaminated time series

MA(1) Coefficient	Standard Error	Sigma^2	AIC	BIC
-0.6641	0.0683	1.971	527.54	533.55

The presence of outliers in the series not only affects the model that is to fit, but also inflates the variance estimate. Without outliers, fitted model of the series was ARIMA(1,0,1) and estimated variance was 1.06, but now selected model is ARIMA(0,1,1) and estimated model variance is 1.971.

Step 2: Obtained residuals from the fitted model are given in Appendix-2.

Step 3: Computed variance estimation is 1.311. This is the MAD for variance estimation which is given in 3rd step of section 3.8 and should not be confused with usual sample standard deviation.

Step 4: All four type of tau statistics for each observation are computed and given in Appendix-3.

Step 5: Maximum tau statistics for each observation is kept and the list of kept tau statistics is given in Appendix-4

Step 6: This step determines possible outliers. Then, selects ones that are notable in respect of rules given 6th step of section 3.8. and ignores the others. Critical value is taken as C=3.5, which corresponds to %95 confidence level, and observations of which absolute value are greater than critical value are kept and shown in Table 4.4.

Table 4.4. Types and locations of detected possible outliers

Time	Type	Estimated Effect	Tau Statistic
15	AO (Additive Outlier)	6.5828	5.5043
33	AO (Additive Outlier)	-5.8776	-4.9147
110	LC (Level Change)	4.6118	4.7051
111	LC (Level Change)	3.8865	3.9651

Now, as mentioned given 6th step of section 3.8. if there are same type of outliers at consecutive time points, the one that has greater tau statistic in absolute value should select. In that regard, LC at time 110 is selected and LC at time 111 is

discarded since $\left|\hat{\lambda}_{111,LC}\right| = \left|3.9651\right| < \left|\hat{\lambda}_{110,LC}\right| \left|4.7051\right|$. Then, the list of outliers that are detected at the end of this step is shown in Table 4.5.

Table 4.5. Types and locations of detected outliers

Time	Type	Estimated Effect	Tau Statistic
15	AO (Additive Outlier)	6.5828	5.5043
33	AO (Additive Outlier)	-5.8776	-4.9147
110	LC (Level Change)	4.6118	4.7051

As noticed, three outliers that we have introduced are detected correctly. An AO at t=15 with estimated effect 6.5828 and tau statistic 5.5043, and an AO at t=33 with estimated effect -5.8776 and tau statistic -4.9147, and also an LC at t=110 with estimated effect 4.6118 and tau statistic 4.7051.

Step 7: The effects of detected outliers are removed from the contaminated time series according to the type of outliers, such

$$\breve{\varepsilon}_t = \begin{cases} \hat{\varepsilon}_t - \hat{\omega}_A \hat{\pi}(B) I_t^{(T)} \ , \ \text{t} = 15, 33 \ \text{ for AO} \\ \hat{\varepsilon}_t - \hat{\omega}_L \hat{l}(B) I_t^{(T)} \ , \ \text{t} \geq 110 \ \text{ for LC.} \end{cases}$$

Adjusted series is given in Appendix-5. The plot of contaminated time series (dotted blue line) and adjusted series (colored black), which is obtained by deleting outliers' effects from the contaminated series, is shown in Figure 4.3.

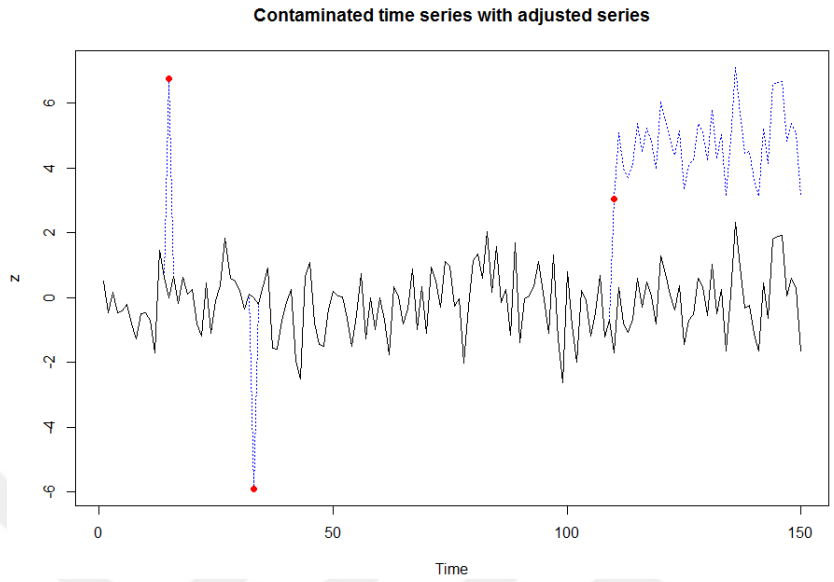


Figure 4.3. Contaminated time series with adjusted series

Step 8: Using the new residuals, which is obtained in step 7, final parameter model is fitted. ARIMA(1,0,1) model is fitted to the adjusted time series by using auto.arima function of RStudio and result is given in Table 4.6.

Table 4.6. Fitted ARIMA to the adjusted time series

	AR(1)	**MA(1)**
Coefficients	0.1226	-0.1301
Standard Error	0.0815	0.0894
Sigma^2 estimated as 0.9375: log likelihood=-207.01		
AIC= 420.01	**AICc**= 420.17	**BIC**= 429.04

It would be appropriate to remark that when series have outliers, fitted model, which is given in Table 4.3, was ARIMA(0,1,1) and estimated variance was 1.971, but now after deleting outliers from the series, fitted model, which is given in Table 4.6, is ARIMA(1,0,1) and estimated variance drops 0.9375. It should be

also remarked that this final model, in Table 4.6, precedes than fitted model to simulated series, in Table 4.1, in respect of both AIC/BIC criterion and small estimated variance.

Step 9: Steps 1 through 8 are repeated to new fitted time series and no more outliers are found.

In this example, we have illustrated how outlier detection procedure can apply to a simulated time series, which is then contaminated by tree outliers, step by step. Outliers that we have introduced are detected correctly. Next, estimated effects of outliers removed from the contaminated time series and obtained adjusted series. Then, we fit ARIMA model to the adjusted series. The results show that existence of outliers in time series causes not only misidentification of the fitted model but also significantly inflate the estimated variance.

To examine the power of the outlier detection procedure, in the example of simulated time series, which is given in Appendix-1, we have changed just the LC in the contaminated series from time 110 to 100 and left same two AO types of outliers at time 15 and 33. The plot of this series is given in Figure 4.4.

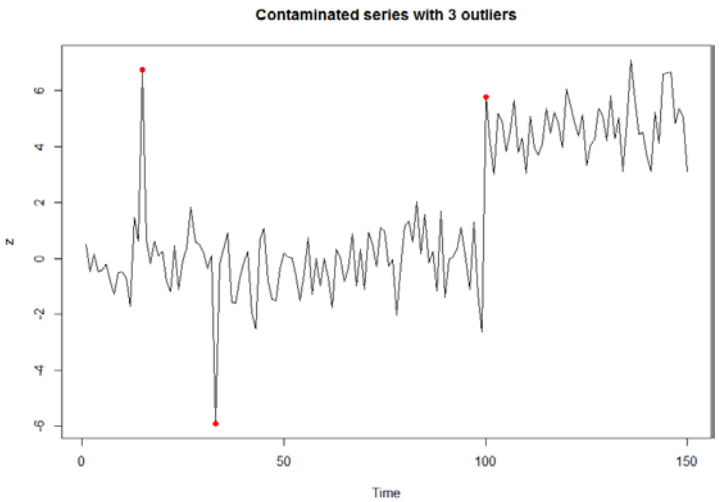


Figure 4.4. Contaminated time series with LC shifted from t=110 to t=100

Then, we run the outlier detection procedure again for this contaminated series and detected outliers are given in Table 4.7.

Table 4.7. Detected outliers for contaminated series that LC shifted t=100

Time	Type	Tau Statistics	Estimated Effect
15	AO	5.8217	6.6762
33	AO	-5.1044	-5.8537
99	AO	-4.2618	-4.8874
100	LC	6.4630	5.4274

As it is expected, procedure detects two AO at time 15 and 33, and an LS at time 100. But surprisingly, procedure detects a new AO at t=99 with estimated effect -4.26 and tau statistic -4.88. This new detected outlier was not determined as an outlier when we carried out outlier detection procedure which is performed before introducing three outliers. This may be because of the effect of level change which occurs right after that point and can be seen as swamping effect of it. As a matter of fact, the procedure checks consecutive outliers. If outliers of same type exceed the critical value in consecutive time points, higher tau statistic in absolute value is kept and the others are removed. If two different types of consecutive outliers exceed the critical value, both are kept. In this context, the procedure should be examined in detail.

To examine the power of the outlier detection procedure, we also contaminated the simulated time series by only adding a TC of size 4 at $t = 105$ till $t = 113$. The plot of this series is given in Figure 4.5.

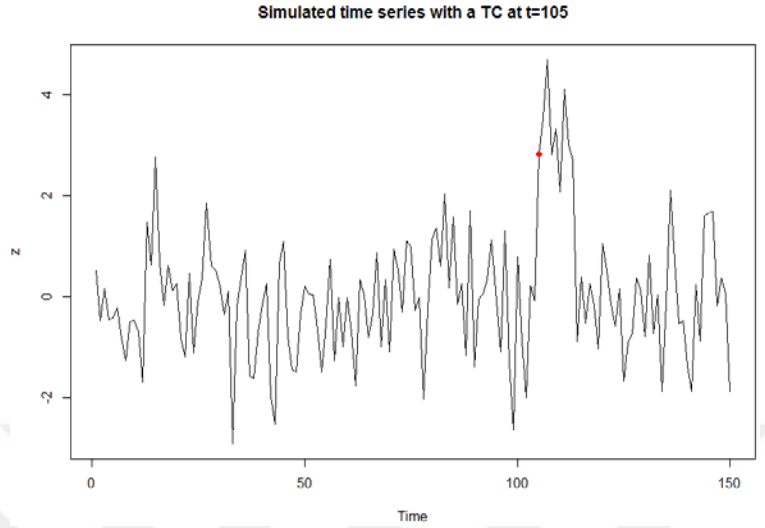


Figure 4.5. Contaminated time series with only TC of size 4 at t=105 till t=113

Then we run the outlier detection procedure. Although outliers can obviously be seen in the plot of contaminated series, procedure did not detect any outlier. But, when we change the size of this outlier from 4 to 5, which is shown in Figure 4.6, procedure correctly detects TC at time 105 with tau statistic 3.6702 and estimated effect 4.2958.

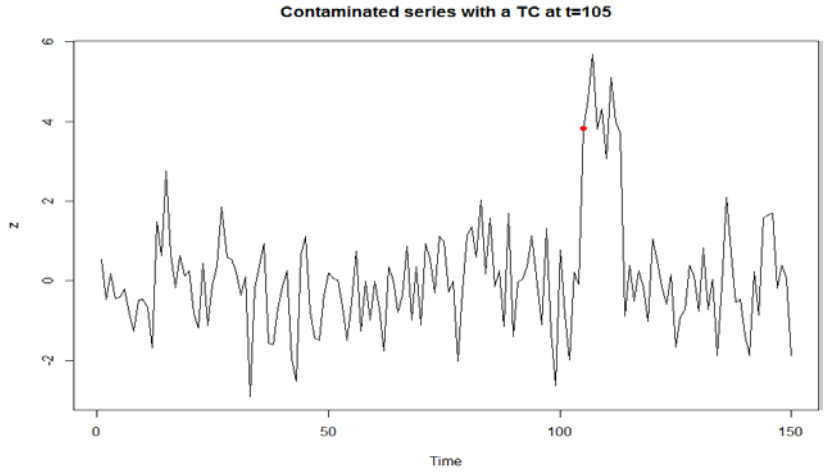


Figure 4.6. Contaminated time series with only TC of size 5 at t=105 till t=113

This result may be seen as an efficiency of the procedure for determining big size (4) of outliers comparing to huge size (5) of outliers. The reason of this may be that TC interval has a value which is smaller than the biggest value of observations in absolute value.

5. CONCLUSIONS

In the first three sections of this thesis, it is presented the theories of that concepts and components of time series, stationarity tests, ARMA models, types of outliers and outlier detection procedure.

In application section, outlier detection procedure is illustrated with simulated time series. Firstly, simulated series is contaminated with three outliers. Secondly, steps of the procedure are implemented to this series. The procedure correctly detected the outliers. Then, effects of these outliers removed from the series and adjusted series is obtained. It is seen that fitted model to adjusted series is significantly different from the fitted model to contaminated series in respect of both small estimated variance and small standard errors of parameters. This situation points out how outlier detection is important in time series. In this context, existence of outliers should be tested at the beginning of analysis. If any outlier is detected in series, the effects of outlier should be removed from the series. This is important in terms of having reliable statistical inferences.

As mentioned at the end of application section, the procedure has difficulty in determining consecutive two different types of outliers. Within this scope, the procedure can be revised and a new step can be added to avoid this confusion. In the presence of two different consecutive outliers, selecting only the one that has higher tau statistic in absolute value may avoid this confusion. Doing this may be seen as has a disadvantage at first glance, just as when real two different types of outliers occur at consecutive time points, procedure will ignore one of them. But, ignored one will be detected thanks to last step of the procedure, which says: run outlier detection procedure for adjusted series. So, by means of this step, ignored one will be identified.

Lastly, as mentioned at the end of application section, procedure also has difficulty in determining some outlier, which can obviously be seen in the plot of the series. This may be seen as an efficiency of the procedure. In this context, there may be made improvement in the outlier detection procedure to determine this kind of outlier rather than the one that has enormous size of outlier.

REFERENCES

Abraham, B. and Yatawara, N., 1988. A Score Test for Detection of Time Series Outliers. *Journal of Time Series Analysis*, 9(2), 109-119.

Abraham, B. and Chuang, A., 1989. Outlier Detection and Time Series Modeling. *Technometrics*, Vol. 31, No. 2 (May, 1989), pp. 241-248

Balke, N. S., 1993. Detecting level shifts in time series. *Journal of Business & Economic Statistics*, 11(1), 81-92.

Beckman, R. and Cook, R.D., 1983. Outliers. *Technometrics*, Vol. 25, No. 2 (May, 1983), pp. 119-149.

Bruce, A. and Martin, R. D., 1989. Leave-k-out Diagnostics for Time Series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(3), 363-401.

Box, J., Jenkins, G. and Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. *ed: Prentice Hall.*

Chang, I., Tiao, G. and Chen, C., 1988. Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*, 30(2), 193-204.

Chen, C. and Liu, L., 1993. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 88(421), 284-297.

_____, 1993. Forecasting Time Series with Outliers. *Journal of Forecasting*, 12(1), 13-35.

Cook, R.D. and Wisberg, S., 1982. Residuals and Influence in Regression. *New York: Chapman and Hall*.

Cryer, J.D. and Chan K.S., 2008. Time Series Analysis with Applications in R. *SpringerLink, Springer eBooks.*

Dickey, D. A. and Fuller, W. A., 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association,* 74(366a), 427-431.

_____, 1981. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica: Journal of the Econometric Society*, 1057-1072.

Fox, A.J., 1972. Outliers in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34, No. 3(1972), pp. 350-363

Fuller, W.A., 1996. Introduction to Statistical Time Series. *John Wiley & Sons.*

Hamilton, J. D., 1994. Time Series Analysis. *Princeton, NJ: Princeton University Press.*

Hodge, V.J. and Austin J., 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85-126, October 2004.

Hui, Y. V. and Lee, A. H., 1992. Assessing Influence in Time Series *Communications in Statistics-Theory and Methods*, 21(12), 3463-3478.

Kaiser, R. and Maravall, A., 1999. Seasonal Outliers in Time Series.

Ljung, G. M. and Box, G. E., 1978. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 65(2), 297-303.

Ljung, G. M., 1993. On Outlier Detection in Time Series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2), 559-567.

Lu, J., Shi, L. and Chen, F., 2012. Outlier Detection in Time Series Models Using Local Influence Method. *Communications in Statistics-Theory and Methods*, 41(12), 2202-2220.

Peña, D., 1990. Influential Observations in Time Series. *Journal of Business & Economic Statistics*, 8(2), 235-241.

Peña, D., Tiao, G. C., and Tsay, R. S., 2011. A Course in Time Series Analysis (Vol. 322). *John Wiley & Sons.*

Sánchez, M.J. and Peña, D., 2003. The identification of multiple outliers in ARIMA models. *Communications in Statistics-Theory and Methods*, 32(6), 1265-1287.

Strickland, J., 2015. Components of Time Series Data. *https://www.linkedin.com/pulse/component-time-series-data-jeffrey-strickland-ph-d-cmsp*

Subba Rao, S., 2018. A Course in Time Series Analysis. *https://www.stat.tamu.edu/~suhasini/teaching673/time_series.pdf p.144-145.*

Tsay, R.S., 1988. Outliers, Level Shifts, and Variance Changes in Time Series. *Journal of Forecasting, Vol. 7, 1-20*

Tsay, R. S., Pena, D. and Pankratz, A. E., 2000. Outliers in Multivariate Time Series. *Biometrika*, 87(4), 789-804.

Yaffee, R. A. and McGee, M., 2000. An Introduction to Time Series Analysis and Forecasting: with applications of SAS and SPSS. Academic Press, INC.

CIRRICULUM VITAE

I was born in 1993 and completed my primary and secondary education in Adana. I graduated from the Department of Statistic at Cukurova University in 2016. During the summers of 2015 and 2016, I joined to “Work and Travel USA” program, in this context, I was able to improve my English and cultural knowledge. In the fall semester of 2016, I was admitted to the MSc program in the same department at the same university. In the second half of 2018, I started to work at OSYM, assessment selection and placement center, as statistician.

APPENDICES

APPENDIX-1: Simulated time series from ARIMA(1,0,1) process with phi parameter $\phi_1 = 0.5$ and theta parameter $\theta_1 = 0.4$, and $\sigma_\varepsilon^2 = 1$, and n=150.

1	0.5277	-0.4567	0.1668	-0.4504	-0.4058
6	-0.2146	-0.8098	-1.2648	-0.4894	-0.4653
11	-0.6796	-1.6949	1.4781	0.6268	**2.7556**
16	0.6673	-0.1711	0.6280	0.1152	0.2452
21	-0.8085	-1.1941	0.4524	-1.1088	-0.0969
26	0.3601	1.8533	0.5969	0.5292	0.2169
31	-0.3488	0.1028	**-2.8857**	-0.1906	0.3665
36	0.9327	-1.5718	-1.6042	-0.7001	-0.1390
41	0.2622	-1.9699	-2.5193	0.6645	1.1009
46	-0.7815	-1.4356	-1.4941	-0.3528	0.1968
51	0.0647	0.0328	-0.6327	-1.4950	-0.5331
56	0.7461	-1.2617	-0.0077	-0.9776	-0.0126
61	-0.6895	-1.7519	0.3367	0.0251	-0.8008
66	-0.3512	0.8764	-0.9836	0.3382	-1.0899
71	0.9401	0.5293	-0.3020	1.1114	0.9944
76	-0.2688	-0.0234	-2.0322	-0.2491	1.1466
81	1.3573	0.5974	2.0413	0.1722	1.5842
86	-0.1388	0.2517	-1.1573	1.7035	-1.3919
91	-0.0431	0.0454	0.3430	1.1345	0.1298
96	-1.0906	1.3239	-1.2785	-2.6333	0.7856
101	-0.8612	-1.9880	0.2227	-0.0812	-1.1769
106	-0.4465	0.6789	-1.1973	-0.6757	**-1.9357**
111	0.0970	-1.0020	-1.2919	-0.8945	0.3884
116	-0.5059	0.2484	-0.1566	-1.0194	1.0599
121	0.5080	-0.1371	-0.5855	0.1545	-1.6725
126	-0.9088	-0.7427	0.3825	0.1214	-0.7771
131	0.8166	-0.7249	0.0436	-1.8689	-0.1122
136	2.0973	0.7148	-0.5427	-0.4728	-1.3608
141	-1.8664	0.2392	-0.8692	1.5863	1.6574
146	1.6933	-0.1842	0.3765	0.0709	-1.8688

APPENDIX-2: Residuals of fitted ARIMA(0,1,1) model to the contaminated time series, which is obtained by adding three outliers to the simulated time series.

1	0.0005	-0.8247	0.1631	-0.5052	-0.2840
6	0.0134	-0.5899	-0.8402	0.2124	0.1610
11	-0.1031	-1.0784	2.4538	0.7796	**6.6477**
16	-1.6751	-1.9525	-0.4967	-0.8399	-0.4278
21	-1.3441	-1.2726	0.7948	-1.0321	0.3245
26	0.6755	1.9386	0.0375	-0.0451	-0.3400
31	-0.7958	-0.0785	**-6.0421**	1.6873	1.6806
36	1.6761	-1.3869	-0.9510	0.2684	0.7382
41	0.8903	-1.6387	-1.6383	2.0920	1.8293
46	-0.6651	-1.1017	-0.7817	0.6209	0.9623
51	0.4991	0.3015	-0.4598	-1.1654	0.1861
56	1.4036	-1.0779	0.5342	-0.6152	0.5614
61	-0.3072	-1.2640	1.2506	0.5205	-0.4843
66	0.1284	1.3152	-0.9865	0.6648	-0.9885
71	1.3735	0.5022	-0.4965	1.0803	0.5974
76	-0.8632	-0.3233	-2.2247	0.3025	1.6009
81	1.2732	0.0856	1.4968	-0.8759	0.8283
86	-1.1699	-0.3870	-1.6670	1.7529	-1.9259
91	0.0710	0.1372	0.3811	1.0431	-0.3073
96	-1.4241	1.4643	-1.6276	-2.4309	1.8056
101	-0.4509	-1.4294	1.2607	0.5372	-0.7432
106	0.2364	1.2870	-1.0253	-0.1609	**3.6331**
111	4.4529	1.8572	0.9434	1.0266	1.9618
116	0.4028	1.0275	0.2724	-0.6791	1.6290
121	0.5319	-0.2968	-0.6471	0.3102	-1.6140
126	-0.3119	-0.0371	1.0953	0.4674	-0.5896
131	1.2085	-0.7374	0.2703	-1.7305	0.6107
136	2.6156	0.3471	-1.0195	-0.6071	-1.2932
141	-1.3688	1.2009	-0.3124	2.2525	1.5659
146	1.0700	-1.1594	-0.2100	-0.4495	-2.2385

APPENDIX-3: All four type of tau statistics for each observation.

	IO	AO	LS	TC		IO	AO	LS	TC
1	0.0004	0.2615	-0.4228	-0.0979	33	**-4.6086**	**-4.9148**	**-2.2933**	**-3.5080**
2	-0.6290	-0.4708	-0.6371	-0.4961	34	1.2870	0.6971	1.7349	1.6852
3	0.1244	0.3259	-0.2512	-0.0672	35	1.2818	1.0427	1.1635	1.4576
4	-0.3854	-0.2095	-0.5183	-0.5401	36	1.2784	1.5653	0.3089	0.6616
5	-0.2166	-0.0836	-0.3466	-0.4861	37	-1.0578	-0.8519	-0.9740	-1.1877
6	0.0102	0.1856	-0.2781	-0.5805	38	-0.7254	-0.8262	-0.2758	-0.5360
7	-0.4500	-0.3525	-0.4303	-1.0823	39	0.2047	0.0335	0.4013	0.3600
8	-0.6409	-0.7930	-0.1413	-1.0657	40	0.5631	0.5427	0.3739	0.4686
9	0.1620	-0.0913	0.5087	-0.4419	41	0.6791	0.9765	-0.0709	-0.0700
10	0.1228	-0.1914	0.5835	-0.5069	42	-1.2499	-1.1792	-0.8713	-1.4305
11	-0.0786	-0.5649	0.7404	-0.4633	43	-1.2496	-1.7751	0.0952	-0.4369
12	-0.8226	-1.8723	1.2034	0.1078	44	1.5956	1.2351	1.5501	1.7945
13	1.8716	0.8811	2.7380	2.7051	45	1.3953	1.5846	0.5378	0.8807
14	0.5946	-0.4272	2.0158	2.6639	46	-0.5073	-0.2272	-0.7610	-0.9009
15	**5.0705**	**5.5044**	**2.3659**	**4.3876**	47	-0.8403	-0.7995	-0.5747	-0.9774
16	-1.2777	-0.4310	-2.1455	-1.2319	48	-0.5962	-0.8686	0.0806	-0.3068
17	-1.4892	-0.9395	-1.7923	-1.1726	49	0.4736	0.1614	0.7925	0.7452
18	-0.3788	0.1105	-1.0223	-0.3951	50	0.7340	0.6008	0.6602	0.8446
19	-0.6406	-0.1932	-1.1128	-0.7149	51	0.3807	0.4193	0.1678	0.3880
20	-0.3263	0.1408	-0.9545	-0.7581	52	0.2299	0.4244	-0.1759	-0.0171
21	-1.0252	-0.7479	-1.0699	-1.2749	53	-0.3507	-0.1585	-0.5237	-0.6027
22	-0.9707	-1.0513	-0.4569	-0.8022	54	-0.8889	-0.9778	-0.3938	-0.6450
23	0.6062	0.5829	0.4048	0.2865	55	0.1419	-0.0566	0.4076	0.4109
24	-0.7873	-1.0363	-0.0729	-0.3849	56	1.0706	1.1902	0.4540	0.6642
25	0.2475	-0.1391	0.7764	0.8621	57	-0.8221	-0.8074	-0.5215	-0.6730
26	0.5152	0.1582	0.8904	1.4212	58	0.4074	0.4731	0.1403	0.1387
27	1.4787	1.5615	0.7607	1.8146	59	-0.4693	-0.4918	-0.2475	-0.4464
28	0.0286	0.3596	-0.5191	0.4647	60	0.4282	0.4921	0.1557	0.0325
29	-0.0344	0.4549	-0.8138	0.1739	61	-0.2343	-0.1690	-0.2477	-0.6240
30	-0.2593	0.3761	-1.1867	-0.3715	62	-0.9641	-1.2568	-0.1092	-0.6613
31	-0.6070	0.0888	-1.4950	-1.0432	63	0.9538	0.7418	0.9209	0.7678
32	-0.0599	0.8852	-1.5678	-1.6113	64	0.3970	0.3522	0.3129	0.0861

APPENDIX-3: All four type of tau statistics for each observation. **(continuous)**

	IO	AO	LS	TC		IO	AO	LS	TC
65	-0.3694	-0.5223	0.0242	-0.3570	97	1.1168	1.6277	-0.1519	0.2188
66	0.0979	-0.1446	0.4523	0.2017	98	-1.2414	-0.7881	-1.4860	-1.9053
67	1.0032	1.0256	0.5708	0.4852	99	-1.8541	-2.0283	-0.8401	-1.6480
68	-0.7525	-0.8670	-0.2698	-0.7044	100	1.3772	1.3842	0.8223	0.4093
69	0.5071	0.4245	0.4408	0.1751	101	-0.3439	-0.2797	-0.3122	-1.3013
70	-0.7539	-1.0929	0.0929	-0.3282	102	-1.0903	-1.4463	-0.0830	-1.4779
71	1.0476	0.8289	0.9887	1.0203	103	0.9616	0.6405	1.1024	-0.1406
72	0.3830	0.3353	0.3093	0.3281	104	0.4098	0.2065	0.5774	-1.0736
73	-0.3787	-0.5414	0.0345	0.0119	105	-0.5669	-1.0305	0.4082	-1.8151
74	0.8240	0.8366	0.4783	0.7548	106	0.1803	-0.5254	1.2528	-1.1889
75	0.4557	0.7539	-0.2074	-0.0617	107	0.9817	0.3096	1.6834	-0.9826
76	-0.6584	-0.3951	-0.8253	-1.1153	108	-0.7820	-1.9563	1.4296	-1.8256
77	-0.2466	-0.0292	-0.5015	-1.0550	109	-0.1227	-2.0401	3.0330	0.0575
78	-1.6969	-2.0360	-0.4775	-1.4673	110	**2.7711**	**0.9029**	**4.7051**	**2.8619**
79	0.2307	-0.4181	1.1911	0.6780	111	3.3964	2.2182	3.9651	2.8583
80	1.2211	0.7307	1.5338	1.5382	112	1.4166	0.6208	2.1470	1.0609
81	0.9711	0.7569	0.9349	1.2019	113	0.7196	-0.0225	1.6382	0.6696
82	0.0653	-0.1045	0.3145	0.6856	114	0.7830	0.0532	1.6566	0.9873
83	1.1417	1.3212	0.4001	1.1218	115	1.4963	1.0599	1.6130	1.3379
84	-0.6681	-0.4964	-0.6827	-0.1976	116	0.3073	-0.0373	0.7443	0.4671
85	0.6318	1.0380	-0.2759	0.3940	117	0.7837	0.5983	0.7749	0.7181
86	-0.8923	-0.5305	-1.1266	-0.8514	118	0.2078	0.1098	0.2845	0.2106
87	-0.2952	0.0215	-0.6918	-0.4935	119	-0.5180	-0.8314	0.1944	0.1512
88	-1.2715	-1.3086	-0.7095	-0.7343	120	1.2425	1.1661	0.8759	1.3488
89	1.3370	1.6123	0.3631	0.7340	121	0.4057	0.6065	-0.0799	0.3379
90	-1.4689	-1.4262	-0.9584	-1.1483	122	-0.2264	0.0451	-0.5770	-0.3437
91	0.0542	-0.0555	0.2105	0.3028	123	-0.4936	-0.2991	-0.6140	-0.5524
92	0.1046	-0.0143	0.2560	0.5081	124	0.2366	0.5526	-0.3689	-0.3817
93	0.2907	0.2341	0.2676	0.7453	125	-1.2310	-1.1837	-0.8218	-1.2982
94	0.7956	1.0460	0.0758	0.7458	126	-0.2379	-0.4183	0.1484	-0.2417
95	-0.2344	0.1604	-0.7815	-0.3598	127	-0.0283	-0.3420	0.4912	0.2246
96	-1.0862	-0.9285	-0.9130	-0.7324	128	0.8355	0.6714	0.7715	0.7868

APPENDIX-3: All four type of tau statistics for each observation. **(continuous)**

	IO	AO	LS	TC
129	0.3565	0.3532	0.2212	0.2092
130	-0.4497	-0.5755	-0.0683	-0.1825
131	0.9217	1.0171	0.4034	0.5235
132	-0.5625	-0.5071	-0.4302	-0.6380
133	0.2061	0.2921	-0.0146	-0.2205
134	-1.3199	-1.6562	-0.2540	-0.7131
135	0.4658	-0.0411	1.1034	1.2383
136	1.9950	2.0385	1.1371	1.8252
137	0.2647	0.6929	-0.5337	-0.1710
138	-0.7776	-0.3883	-1.1016	-1.1896
139	-0.4630	-0.1526	-0.7834	-1.1708
140	-0.9863	-0.9486	-0.6583	-1.4664
141	-1.0441	-1.5076	0.1191	-0.8000
142	0.9160	0.4223	1.3551	0.9267
143	-0.2383	-0.9500	1.0094	0.7477
144	1.7181	1.2578	1.7899	2.4059
145	1.1944	1.1752	0.7595	1.7020
146	0.8161	1.2515	-0.2069	0.7816
147	-0.8843	-0.4604	-1.2515	-0.6890
148	-0.1602	0.3115	-0.8921	-0.3187
149	-0.3428	0.2187	-1.2302	-0.9063
150	-1.7074	-1.7074	-1.7074	-1.7074

APPENDIX-4: Maximum tau statistics with its type for each observation.

	MaxAbs TauStats	Type
1	0.4228	"LS"
2	0.6371	"LS"
3	0.3259	"AO"
4	0.5401	"TC"
5	0.4861	"TC"
6	0.5805	"TC"
7	1.0823	"TC"
8	1.0657	"TC"
9	0.5087	"LS"
10	0.5835	"LS"
11	0.7404	"LS"
12	1.8723	"AO"
13	2.7380	"LS"
14	2.6639	"TC"
15	5.5044	"AO"
16	2.1455	"LS"
17	1.7923	"LS"
18	1.0223	"LS"
19	1.1128	"LS"
20	0.9545	"LS"
21	1.2749	"TC"
22	1.0513	"AO"
23	0.6062	"IO"
24	1.0363	"AO"
25	0.8621	"TC"
26	1.4212	"TC"
27	1.8146	"TC"
28	0.5191	"LS"
29	0.8138	"LS"
30	1.1867	"LS"
31	1.4950	"LS"
32	1.6113	"TC"
33	4.9148	"AO"
34	1.7349	"LS"
35	1.4576	"TC"
36	1.5653	"AO"
37	1.1877	"TC"
38	0.8262	"AO"
39	0.4013	"LS"
40	0.5631	"IO"
41	0.9765	"AO"
42	1.4305	"TC"
43	1.7751	"AO"
44	1.7945	"TC"
45	1.5846	"AO"
46	0.9009	"TC"
47	0.9774	"TC"
48	0.8686	"AO"
49	0.7925	"LS"
50	0.8446	"TC"
51	0.4193	"AO"
52	0.4244	"AO"
53	0.6027	"TC"
54	0.9778	"AO"
55	0.4109	"TC"
56	1.1902	"AO"
57	0.8221	"IO"
58	0.4731	"AO"
59	0.4918	"AO"
60	0.4921	"AO"
61	0.6240	"TC"
62	1.2568	"AO"
63	0.9538	"IO"
64	0.3970	"IO"
65	0.5223	"AO"
66	0.4523	"LS"
67	1.0256	"AO"
68	0.8670	"AO"
69	0.5071	"IO"
70	1.0929	"AO"
71	1.0476	"IO"
72	0.3830	"IO"
73	0.5414	"AO"
74	0.8366	"AO"
75	0.7539	"AO"
76	1.1153	"TC"
77	1.0550	"TC"
78	2.0360	"AO"
79	1.1911	"LS"
80	1.5382	"TC"
81	1.2019	"TC"
82	0.6856	"TC"
83	1.3212	"AO"
84	0.6827	"LS"
85	1.0380	"AO"
86	1.1266	"LS"
87	0.6918	"LS"
88	1.3086	"AO"
89	1.6123	"AO"
90	1.4689	"IO"
91	0.3028	"TC"
92	0.5081	"TC"
93	0.7453	"TC"

APPENDIX-4: Maximum tau statistics with its type for each observation **(continuous).**

	MaxAbs TauStats	Type
94	1.0460	"AO"
95	0.7815	"LS"
96	1.0862	"IO"
97	1.6277	"AO"
98	1.9053	"TC"
99	2.0283	"AO"
100	1.3842	"AO"
101	1.3013	"TC"
102	1.4779	"TC"
103	1.1024	"LS"
104	1.0736	"TC"
105	1.8151	"TC"
106	1.2528	"LS"
107	1.6834	"LS"
108	1.9563	"AO"
109	3.0330	"LS"
110	4.7051	"LS"
111	3.9651	"LS"
112	2.1470	"LS"
113	1.6382	"LS"
114	1.6566	"LS"
115	1.6130	"LS"
116	0.7443	"LS"
117	0.7837	"IO"
118	0.2845	"LS"
119	0.8314	"AO"
120	1.3488	"TC"
121	0.6065	"AO"
122	0.5770	"LS"
123	0.6140	"LS"
124	0.5526	"AO"
125	1.2982	"TC"
126	0.4183	"AO"
127	0.4912	"LS"
128	0.8355	"IO"
129	0.3565	"IO"
130	0.5755	"AO"
131	1.0171	"AO"
132	0.6380	"TC"
133	0.2921	"AO"
134	1.6562	"AO"
135	1.2383	"TC"
136	2.0385	"AO"
137	0.6929	"AO"
138	1.1896	"TC"
139	1.1708	"TC"
140	1.4664	"TC"
141	1.5076	"AO"
142	1.3551	"LS"
143	1.0094	"LS"
144	2.4059	"TC"
145	1.7020	"TC"
146	1.2515	"AO"
147	1.2515	"LS"
148	0.8921	"LS"
149	1.2302	"LS"
150	1.7074	"IO"

APPENDIX-5: Adjusted time series.

1	0.5300	33	**0.0000**	65	-0.8000	97	1.3200	129	0.3454
2	-0.4600	34	-0.1900	66	-0.3500	98	-1.2800	130	-0.5546
3	0.1700	35	0.3700	67	0.8800	99	-2.6300	131	1.0454
4	-0.4500	36	0.9300	68	-0.9800	100	0.7900	132	-0.4946
5	-0.4100	37	-1.5700	69	0.3400	101	-0.8600	133	0.2654
6	-0.2100	38	-1.6000	70	-1.0900	102	-1.9900	134	-1.6446
7	-0.8100	39	-0.7000	71	0.9400	103	0.2200	135	0.1154
8	-1.2600	40	-0.1400	72	0.5300	104	-0.0800	136	2.3254
9	-0.4900	41	0.2600	73	-0.3000	105	-1.1800	137	0.9354
10	-0.4700	42	-1.9700	74	1.1100	106	-0.4500	138	-0.3146
11	-0.6800	43	-2.5200	75	0.9900	107	0.6800	139	-0.2446
12	-1.6900	44	0.6600	76	-0.2700	108	-1.2000	140	-1.1346
13	1.4800	45	1.1000	77	-0.0200	109	-0.6800	141	-1.6446
14	0.6300	46	-0.7800	78	-2.0300	110	**-1.7146**	142	0.4654
15	**0.0000**	47	-1.4400	79	-0.2500	111	0.3254	143	-0.6446
16	0.6700	48	-1.4900	80	1.1500	112	-0.7746	144	1.8154
17	-0.1700	49	-0.3500	81	1.3600	113	-1.0646	145	1.8854
18	0.6300	50	0.2000	82	0.6000	114	-0.6646	146	1.9154
19	0.1200	51	0.0600	83	2.0400	115	0.6154	147	0.0454
20	0.2500	52	0.0300	84	0.1700	116	-0.2846	148	0.6054
21	-0.8100	53	-0.6300	85	1.5800	117	0.4754	149	0.2954
22	-1.1900	54	-1.4900	86	-0.1400	118	0.0654	150	-1.6446
23	0.4500	55	-0.5300	87	0.2500	119	-0.7946		
24	-1.1100	56	0.7500	88	-1.1600	120	1.2854		
25	-0.1000	57	-1.2600	89	1.7000	121	0.7354		
26	0.3600	58	-0.0100	90	-1.3900	122	0.0854		
27	1.8500	59	-0.9800	91	-0.0400	123	-0.3646		
28	0.6000	60	-0.0100	92	0.0500	124	0.3754		
29	0.5300	61	-0.6900	93	0.3400	125	-1.4446		
30	0.2200	62	-1.7500	94	1.1300	126	-0.6846		
31	-0.3500	63	0.3400	95	0.1300	127	-0.5146		
32	0.1000	64	0.0300	96	-1.0900	128	0.6054		