ABSTRACT

A Study of Time Series Using Univariate Box-Jenkins Model

Yuelin Lu

Director: Jeanne S. Hill

This thesis looks at the Univariate Box-Jenkins Models for time series analysis. It introduces readers to the concepts of time series and this analytical model. Starting from underlying principles of time series, we cover the practice of ARIMA modeling, including a three-stage procedure: identification, estimation and forecasting.

APPROVED BY DIRECTOR OF HONORS THESIS:

____

Dr. Jeanne S. Hill

APPROVED BY THE HONORS PROGRAM:

____

Dr. Andrew Wisely, Director

DATE: ___________________________

A STUDY OF TIME SERIES USING UNIVARIATE BOX-JENKINS MODELS

A Thesis Submitted to the Faculty of
Baylor University
In Partial Fulfillment of the Requirements for the
Honors Program

By
Yuelin Lu

Waco, TX
May 2014

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER ONE

We always try to explore our understanding and realization about the future based on what we know about the present. Though many approaches have successfully fulfilled the prediction task, one that is widely used in a variety fields of study and that is, time series.

Time-series data refers to observations on a variable that occur in a time sequence. By applying techniques and analysis on these data, we strive to find the interrelationship between the data and use it to predict the possible future outcomes, changes or trends. This thesis focuses on using one of the models in time series analysis --- the Univariate Box-Jenkins (UBJ) models to forecast time series data. This thesis will discuss details of this methodology.

As stated above, the field, time series analysis, consists of the techniques which when applied to time series lead to improved knowledge about this data set (Brillinger). It serves to summarize, predict, and describe the behavior of the target data set. According to Brillinger, there are two sides of this field: one is a theoretical side and the other is an applied side. In this thesis, both will be considered. The theoretical side includes the theory of stochastic processes, such as representation, prediction, information, and limit theorems, while applications involve extensions of techniques of statistics like regression, analysis of variance, multivariate analysis and sampling (Brillinger). The theories and techniques involved in analyzing a time series data set will be presented clearly for UBJ models.

An introduction to UBJ models

The UBJ models are named after its two individuals, George E. P. Box and Gwilym M. Jenkins. George E. P. Box was a statistician. He worked in the areas of quality control, time-series analysis, design of experiments and Bayesian inference. His name appears in such statistical results as Box-Jenkins models, Box-Cox transformations, and Box-Behnken designs, to name a few. One of his famous comments would be that "essentially, all models are wrong, but some are useful", where he wrote it in his book on response surface methodology with Norman R. Draper.

Box was born in Gravesend, Kent, England in 1919. He did not begin his journey of statistics until he performed experiments of exposing small animals to poison gas during World War II when he was serving British Army. After the war, he finished his PhD in statistics at the University of London. Later he worked as a statistician for Imperial Chemical Industries (ICI). He was also the founder of the Department of Statistics in the University of Wisconsin-Madison. He retired in 1992, becoming an Emeritus Professor and died on March 28, 2013.

Gwilym Meirion Jenkins was a Welsh statistician and systems engineer born in Gowerton, Swansea, Wales August 12, 1932. He was most known for his pioneering work with George Box on autoregressive moving average models, also called Box-Jenkins models, in time series analysis. He died in 1982 from Hodgkin's lymphoma. G.E.P. Box wrote his obituary.

Returning to the models that bear their names, we not that UBJ models are for univariate or single-series data meaning only past data observations of a single variable of interest. The models are also referred to as auto-regressive integrated moving average (ARIMA) models. We will discuss the two parts of our models, autoregressive (AR) and moving average (MA). This model is used to forecast time-series data, or observations on a variable that occur in a time period.

Though the UBJ models are practical in most of the cases, there are still some preferred conditions for which they are optimal. The first one is for short-term forecasting. As the UBJ models are weighted toward the most recent past, a distant past will make the model less reliable. The second concerns data type. UBJ models will work with either discrete or continuous data. The constraint on both discrete and continuous data is that it should be measured at equally spaced and discrete time intervals. UBJ-ARIMA models are particularly useful for forecasting data series that contain seasonal (or other periodic) variation, including those with shifting seasonal patterns (Pankratz). Another requirement is an adequate sample size. Generally about 50 observations are needed. With regard to seasonal data, a larger sample size may be required. Lastly, UBJ-ARIMA applies only to stationary data: meaning that the mean, variance, and autocorrelation function are unchanged through time.

Time- series data is encountered in areas such as economics, meteorology, sociology and geography ... Some examples include weekly share prices, monthly profits, daily rainfall, wind speed, temperature, and crime figures. The following graphs give some examples of time series plots.

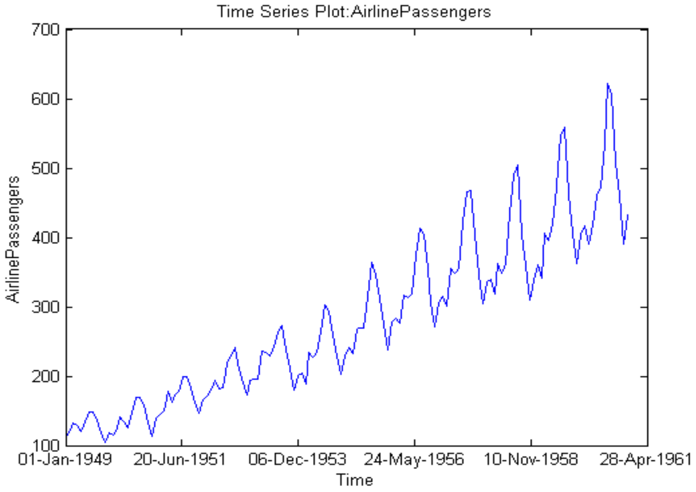


Figure One: a time series data counted monthly of airline passengers from January 1949 to December 1960 (measured in thousand).

There is a clear upper trend in this plot, implying that the latter data are related to the previous ones.

The general procedures for a UBJ modeling process include three stages: identification, estimation, and diagnostic checking. In the first stage, we choose one or more of the most suitable ARIMA models as candidates based on the behaviors of the data. Second, we estimate the parameters of the model(s) chosen at stage one. And lastly, we check the candidate model(s) for adequacy. If the model is satisfactory, then we use it to forecast; otherwise, repeat from stage one. More details of each stage will be explained and illustrated with our analysis of a particular case study later. You have not included the case study, so you need to add that before the conclusion.

There are three advantages to UBJ models. First, they are derived foundational results in classical probability theory and mathematical statistics." Others are often based on ad hoc or intuitive methods. Second, ARIMA model is a family of models instead of a single model. Box and Jenkins provided analysts a guide for how to choose proper models from this family. Third, no other standard single-series model can give forecasts with a smaller mean-squared forecast error than an appropriate ARIMA model.

Now that we have considered the overall picture of the procedures of a UBJ modeling, we need to examine more closely the actual tools used in these procedures. The estimated autocorrelation function (acf) and the estimated partial autocorrelation function (pacf) are the two one of the most essential tools.

CHAPTER TWO

Theoretical analysis

Adjustment of a set of data

We begin our analysis by adjusting the observed data set. There are two steps for us to adjust a set of data if it does not satisfy our requirement of stationarity. The first one is differencing, which is used when the mean of a series is changing over time (i.e. non-stationary). This operation involves calculating successive changes in the values of a data series (z_t). The specific procedures are as the following: we name the first order of differences w_t, where

$$w_t = z_t - z_{t-1}, t = 2, 3, \dots, n \qquad (1)$$

Notice that n starts at 2 instead of 1, and with this we lose one observation. After the differencing, w_t tends to have a more constant mean. If not, we need to redefine w_t to be the following:

$$w_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2}), t = 3, 4, 5, \dots, n \qquad (2)$$

and so forth. A question may be raised here that would this differencing operation change our set of data and thus lead to unreliability? The answer is no, because z_t and w_t are related by definition. Therefore, by appropriate differencing frequently, we can transform a non-stationary series into stationary one, with mean zero.

The second step is expressing the data in deviations from the mean. This tool is used when the mean of a series is constant (stationary), and we consider the mean as a fixed and non-stochastic component of the series. We define a new series $\tilde{z}_t = z_t - \bar{z}$. In this case, the mean of the new series is zero, which is the only difference between $\tilde{z}_t$ and w_t.

Two analytical tools

After adjusting our data set to meet the requirement, we can embark on finding the correlations among each order pairs derived from our data set. Here we introduce two new concepts, the estimated autocorrelation functions (acf) and estimated partial autocorrelation functions (pacf).

First, let us take a look at the estimated autocorrelation functions (acf). It is a collection of the correlation coefficients between each set of ordered pairs $(\tilde{z}_t, \tilde{z}_{t+k})$ and we use r_k to denote the estimated autocorrelation coefficient of observations separated by k time periods within a time series. The true, underlying autocorrelation coefficients however are designated ρ_k, which are estimated by the r_k's statistics. The estimated autocorrelation functions measure the direction and strength of the statistical relationship between ordered pairs of observations of two random variables that are measured k time units apart. The calculation of r_k is given as the following:

$$r_k = \frac{\sum_{t=1}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2} \qquad (3)$$

The maximum number of useful estimated autocorrelations is roughly ¼ of the total observations, which makes it more convenience for calculation. Here is an example of output displaying the estimated autocorrelation function of simulated time-series data.

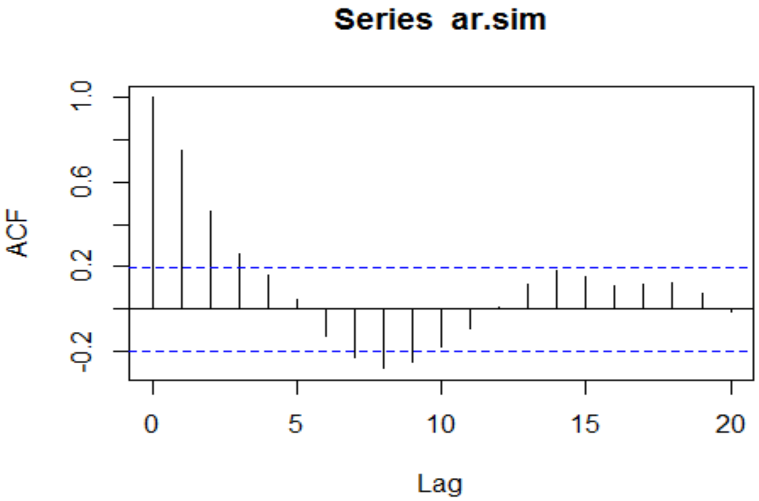


Figure Two: ACF of simulated data of AR (2) model using R

We use the term lag k to represents the number of time periods separating the ordered pairs used to calculate each r_k. Why is the acf so important that we spend this much effort finding it? It is because that the acf is a key element at the identification stage of the UBJ method. The analyst must make a judgment about what ARIMA model(s) might fit the data by examining the patterns in the estimated acf.

Next we consider the partial autocorrelation functions (pacf). It is also used as a guide for choosing the suitable models for our data set. It provides us a way to measure how $\tilde{z}_t$ and $\tilde{z}_{t+k}$ are related and meanwhile takes the effect of the $\tilde{z}$'s intervene into account. For example, if we want to examine the relationship between ($\tilde{z}_t$,$\tilde{z}_{t+k}$), the effect of $\tilde{z}_{t+1}, \tilde{z}_{t+2}, \ldots, \tilde{z}_{t+k}$ needs to be taken into consideration. The estimated partial autocorrelation coefficient is denoted by $\widehat{\square}_{kk}$. As, r_k is the estimated coefficient of ρ_k, $\widehat{\square}_{kk}$ estimates the true partial autocorrelation coefficient $\square_{kk}$. An algorithm for calculating $\square_{kk}$ using least squares is given below. :

(1) $\tilde{z}_{t+1} = \square_{11}\tilde{z}_t + u_{t+1}$, where $\tilde{z}_{t+1}$ and $\tilde{z}_t$ are all the possible ordered pairs of observations; $\square_{11}$ is the true coefficient we want to estimate; u_{t+1} is the error term.

(2) $\tilde{z}_{t+2} = \square_{21}\tilde{z}_{t+1} + \square_{22}\tilde{z}_t + u_{t+2}$, for finding $\square_{22}$.

(3) $\tilde{z}_{t+3} = \square_{31}\tilde{z}_{t+2} + \square_{32}\tilde{z}_{t+1} + \square_{33}\tilde{z}_t + u_{t+3}$, for finding $\square_{33}$ and so on. Then we get

(4) $\widehat{\square}_{11} = r_1$, and

$$\widehat{\square}_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \widehat{\square}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \widehat{\square}_{k-1,j} r_j} \qquad (k = 2, 3, \ldots) \qquad (4)$$

where $\widehat{\square}_{kj} = \widehat{\square}_{k-1,j} - \widehat{\square}_{kk}\widehat{\square}_{k-1,j-1} \qquad (k = 3, 4, \ldots; j = 1, 2, \ldots, k-1)$

Here is a graphical example of pacf function computed from simulated data

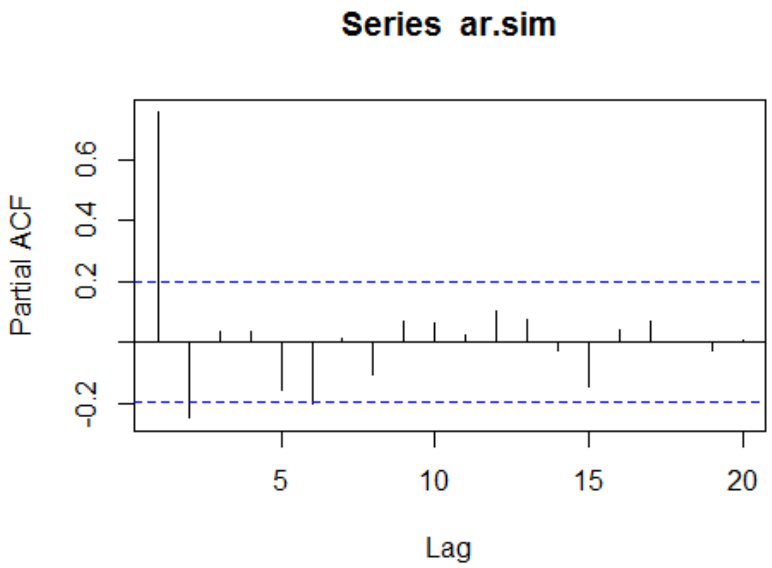


Figure Three: PACF of simulated data of AR (2) model using R

We can use the estimated acf graph to help us decide whether a data series is stationary or not. For a mean-stationary data set, the estimated acf drops off rapidly to zero. Otherwise, it drops slowly to zero. We will examine stationarity more carefully later in Chapter Three, but the following gives a good example of a non-stationary series acf and pacf:

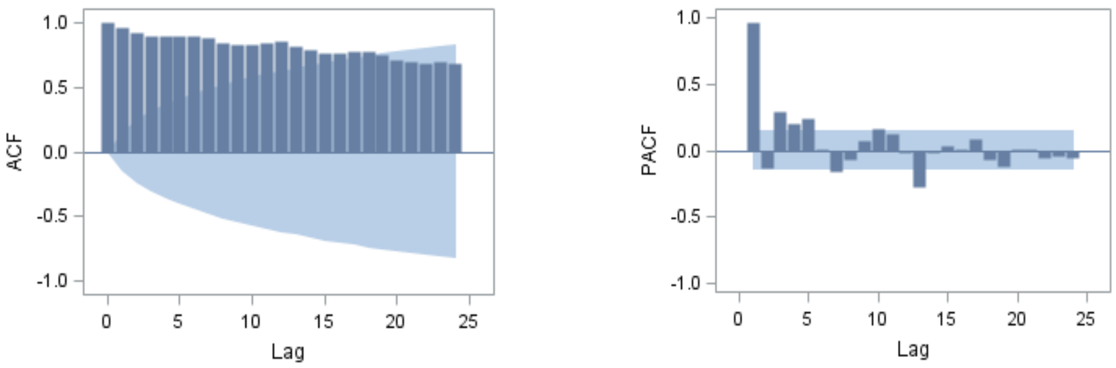


Figure Four: ACF and PACF for non-stationary time series data

So far, we have examined the difference between acf and pacf, how the estimated coefficients are calculated, how to interpret the estimated coefficients, and how the estimated coefficients are represented graphically.

Underlying Statistically Principles

In this section, we are going to discuss the three-stage UBJ procedure, terminology and principles of the UBJ method.

This three-stage process is known as the Process, Realization and Model process. Stage one is process, where all the possible outcomes and a stochastic generating mechanism specify how these observations are related through time. Stage two, which is realization, we construct an acf and a pacf using available data. The last stage, model, is a representation of the process developed by analyzing the realization. Usually, the process stage is the starting place. It gives rise to a realization. After having a realization we choose a model to fit the realization best. Hopefully the model we choose is a good representation of the unknown, underlying process.

There are two common processes we can use to examine the series on stage one: AR process and MA process. AR is the abbreviation for autoregressive process. It associates z_t immediately with the past value z_{t-1}. A general form for AR order one would be

$$z_t = C + \square_1 z_{t-1} + a_t \qquad (5)$$

where *C* is constant term related to the mean of the process; $\square_1$ is a parameter of fixed numerical value; $z_{t-1}$is the one period previous data of the same series; a_t is a random shock element. (Note: the AR order, denoted by AR (n), is the longest time lag associated with a *z* term on the right-hand-side). Now consider MA process. MA is the abbreviation for moving average. It associates z_t with a_{t-1}. Alternatively speaking, an MA term represents a relationship between z_t and a component of a past *z* term, where the component is the appropriately lagged random shock. A general form for MA order one is

$$z_t = C - \theta_1 a_{t-1} + a_t \qquad (6)$$

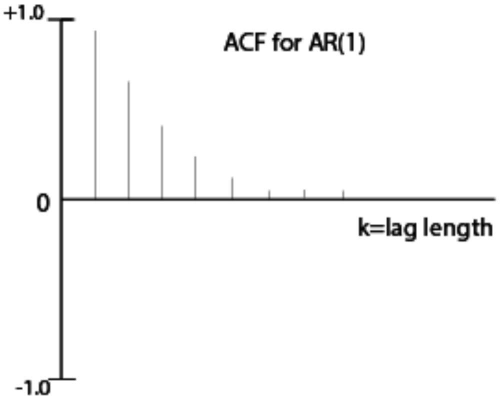
where θ_1 is a fixed value; a_{t-1} is the past random shock; a_t is the present random shock.

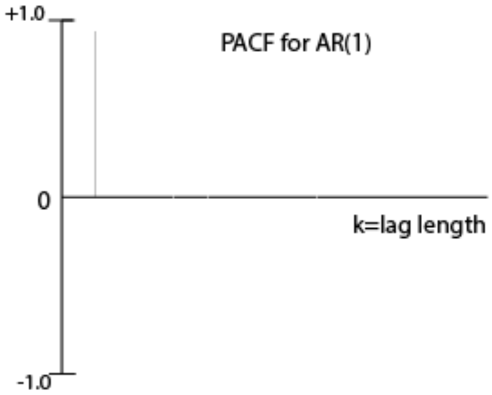
The a_t terms are usually assumed to be normally, identically and independently distributed random variables with a mean of zero and a constant variance .We often refer to these random variables as “white noise.” Any specific value of an observation is composed by three parts: a deterministic part denoted by C, another part reflecting the past data AR and MA, and a pure random term a_t, the presence of this white noise is vital with regard to our consideration that z_t should be at least partly affected by random chance.

Theoretical ACF's and PACF's Vs estimated ACF's and PACF's

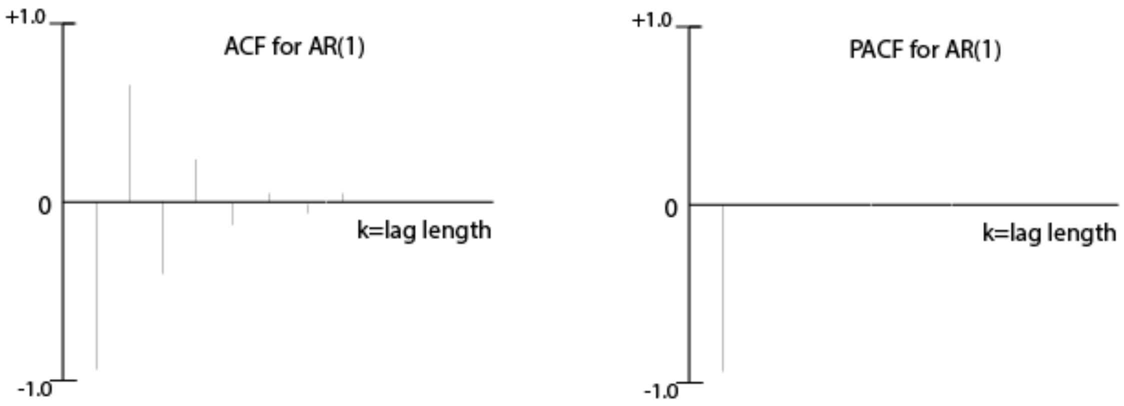
In order for us to distinguish which process a case falls into either AR or MA, it's important to know some characteristics of the theoretical acf's and pacf's. For stationary autoregressive processes (AR), they have theoretical acf that decay or damp out toward zero. However, their theoretical pacf's cut off sharply to zero after a few spikes. The lag length of the last pacf spike equals the AR order (p) of the process. On the other hand, for moving-average processes (MA), the theoretical acf's cut off sharply to zero after a certain number of spikes. The lag length of the last acf spike equals the MA order (q) of the process. The theoretical pacf's decays toward zero. The following graphs are some examples for theoretical acf's and pacf's for stationary AR (1) and MA (1) processes.

Example I: $\square_1$ *is positive*

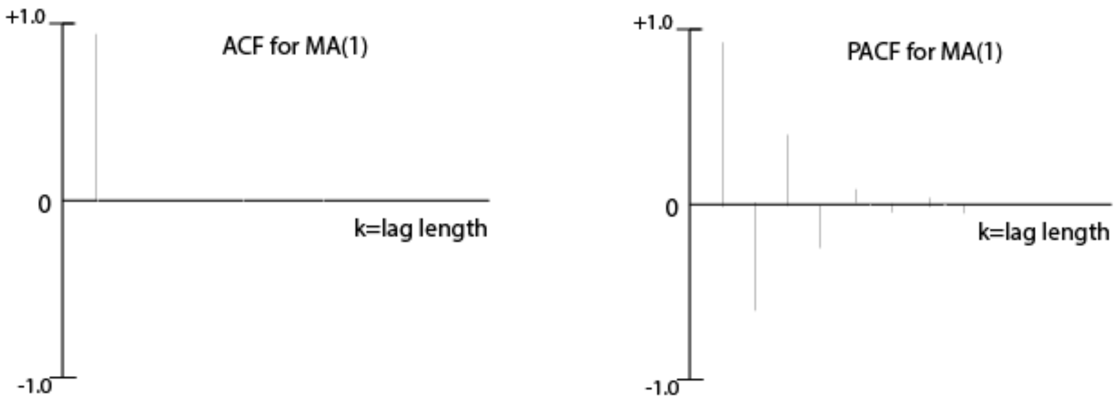




Example II: $\square_1$ is negative



Example I: θ_1 is negative



Example II: θ_1 is positive

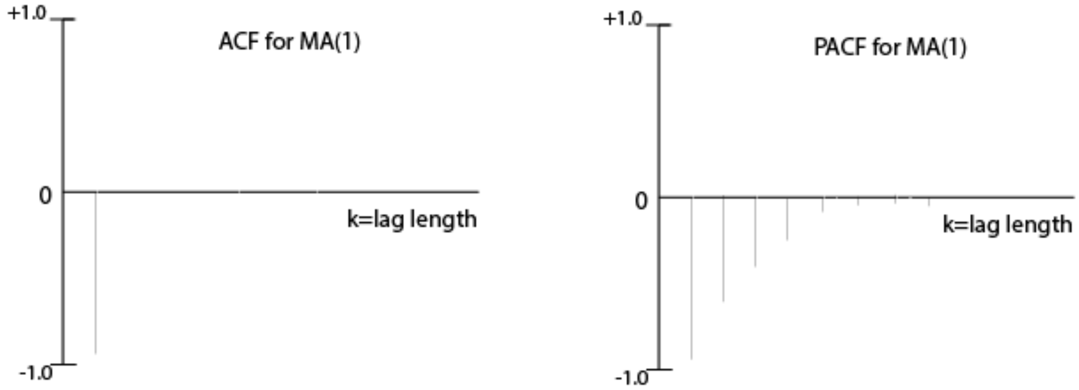


Figure Five: Theoretical ACF and PACF's for AR (1) and MA (1)

The estimated acf and pacf are calculated by using formulas (3) and (4) and the observed data. Next we compare the estimated acf and pacf to a range of common theoretical acf's and pacf's and find a match and select an initial model. During this process, we cannot expect the estimated acf and pacf match exactly to one of the theoretical ones. After all, the observed data are realization based and therefore involve sampling error. Here is an example of estimated acf and pacf for AR (1) process $z_t = 0.5z_{t-1} + a_t$, with $a_t \sim iid\ N(0, \square^2)$. We used the simulated data generated from R.

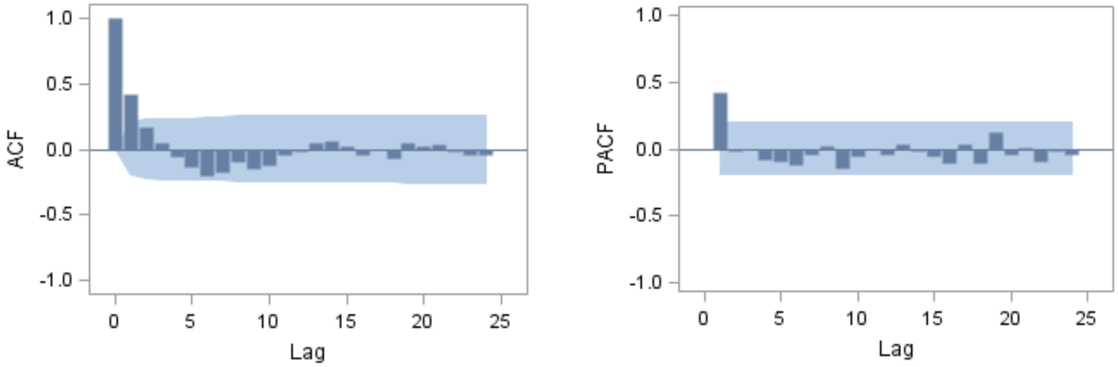


Figure Six: Estimated ACF and PACF for $z_t = 0.5z_{t-1} + a_t$

Standards for a good ARIMA model

Before we introduce more of ARIMA modeling, there are several criteria for a good model that should be kept in mind. The model should be parsimonious or as simplest model that still explains the data. It should also be stationary and invertible. The estimated coefficients should be statistically significant and not highly correlated with each other. The residuals should be uncorrelated and most importantly, the model must forecast well.

Notation and interpretation of ARIMA model

Now we introduce notation and the interpretation of ARIMA models. It is useful to explain the intuition behind the model. As mentioned above, the two common ARIMA processes, the AR (1) and MA (1) are:

$$z_t = C + \square_1 z_{t-1} + a_t$$

$$z_t = C - \theta_1 a_{t-1} + a_t$$

here are three additional common processes researchers use a lot shown as the following:

$$z_t = C + \square_1 z_{t-1} + \square_2 z_{t-2} + a_t \qquad (7)$$

$$z_t = C - \theta_1 a_{t-1} - \theta_2 a_{t-2} + a_t \qquad (8)$$

$$z_t = C + \square_1 z_{t-1} - \theta_1 a_{t-1} + a_t \qquad (9)$$

Equation (7) is called an AR (2) because it contains only AR terms, which is z_t's and the maximum time lag on the AR terms is two (i.e. z_{t-2}). Similarly, equation (8) is MA (2). Equation (9) is more interesting being a mixture of both AR and MA. It is an ARMA (1, 1) process because the order for AR and MA are both one.

The notation ARIMA (*p, d, q*) is used to denote a complete ARIMA process. The p indicates the number of the AR order and q is the number of the MA order. Sometimes we use AR and MA process interchangeably.   It can be shown that an MA (1) process is equivalent to an AR process of infinitely high order and an AR (1) process is equivalent to an MA process of infinitely high order. The *d* represents how many times a series has been differenced. If a series has been differenced, then the letter "I" is added, as well. We can have ARMA (*p, q*) models or ARIMA (*p, d, q*) models.

Another useful notation is the *backshift operator B*. Here is how it operates: $Bz_t = z_{t-1}$. It would be really tempting to think *B* is just a number; however, this algebraic term does not stand number, but it is an operator acting onz_t or a_t. *B* is meaningful. Its task is to shift the time subscripts. For example, B^2gives $B^2 z_t = z_{t-2}$ and more generally, $B^k z_t = z_{t-k}$. The next operator is the *differencing operator* (*1-B*). Recall that a non-stationary series must be differenced before we run analysis on it. Therefore, $(1 - \mathrm{B})z_t = z_t - z_{t-1}$ produces the first differences of z_t. And similarly, $(1 - B)^2 z_t = z_t - 2z_{t-1} + z_{t-2}$ generates the second differences.

CHAPTER THREE

Identification Stage

Theoretical ACF's and PACF's for five common processes

In this chapter, we will introduce the method used to identify a model by matching the estimated acf's and pacf's with the theoretical ones. Although there are numerous ARIMA models proposed, only a few commonly appear in practice. Besides, the characteristics of the rare models are included in the ordinary ones. Knowing the theoretical acf's and pacf's for five common processes is a start.

Below are the models for five common processes.

$$\text{AR (1): } \left(1-\square_1 B\right)\tilde{z}_t = a_t \qquad (10)$$

$$\text{AR (2): } \left(1-\square_1 B-\square_2 B^2\right)\tilde{z}_t = a_t \qquad (11)$$

$$\text{MA (1): } \tilde{z}_t = (1-\theta_1 B)a_t \qquad (12)$$

$$\text{MA (2): } \tilde{z}_t = (1-\theta_1 B-\theta_2 B^2)a_t \qquad (13)$$

$$\text{ARMA (1, 1): } \left(1-\square_1 B\right)\tilde{z}_t = (1-\theta_1 B)a_t \qquad (14)$$

The tables below summarize the characteristics of acf's and pacf's for the common models. (Pankratz, page 122-123)

Process	acf	Pacf
AR	Tails off toward zero (exponential decay or damped sine wave)	Cuts off to zero (after lag p)
MA	Cuts off to zero (after lag q)	Tails off toward zero (exponential decay or damped sine wave)
ARMA	Tails off toward zero	Tails off toward zero

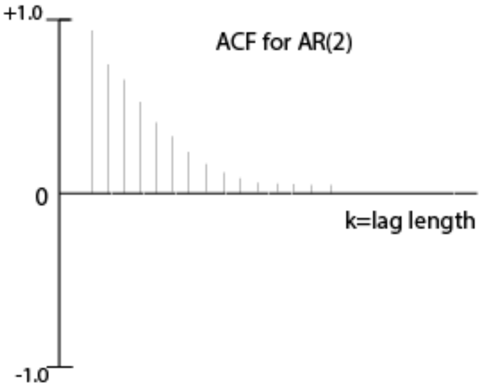
More specifically, with different signs of coefficient in front, each individual may behave differently.

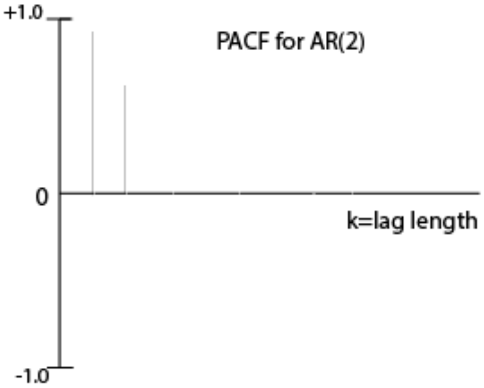
Process	acf	Pacf
AR (1)	Exponential decay: (i) on the positive side if $\phi_1 > 0$; (ii) alternating the sign starting on the negative side if $\phi_1 < 0$.	Spike at lag 1, then cuts off to zero: (i) spike is positive if $\phi_1 > 0$; (ii) spike is negative if $\phi_1 < 0$.
AR (2)	A mixture of exponential decays or a damped sine wave. The exact pattern depends on the signs and	Spikes at lags 1 and 2, and then cuts off to zero.

	sizes of $\square_1$and $\square_2$.	
MA (1)	Spike at lag 1, then cuts off to zero: (i) spike is positive if $\theta_1 < 0$; (ii) spike is negative if $\theta_1 > 0$.	Damps out exponentially: (i) alternating in sign, starting on the positive side, if $\theta_1 < 0$; (ii) on the negative side, if $\theta_1 > 0$.
MA (2)	Spikes at lags 1 and 2, and then cuts off to zero.	A mixture of exponential decays or a damped sine wave. The exact pattern depends on the signs and sizes of $\theta_1$and θ_2.
ARMA (1, 1)	Exponential decay from lag 1: (i) sign of $\square_1$= sign of $(\square_1 - \theta_1)$; (ii) all one sign if $\square_1 > 0$; (iii) alternating in sign if $\square_1 < 0$.	Exponential decay from lag 1: (i) $\square_{11} = \square_1$; (ii) all one sign if $\theta_1 > 0$; (iii) alternating in sign if $\theta_1 < 0$.

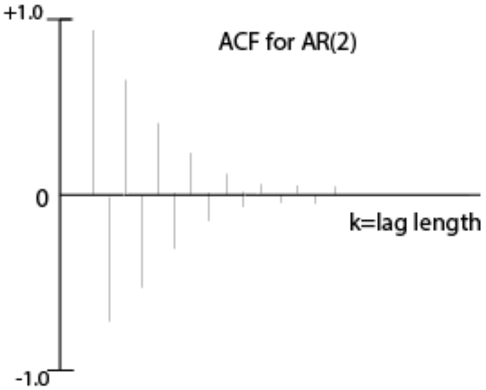
The graphs provided below would provide a better understanding of those shapes. The first examples are for AR (2).

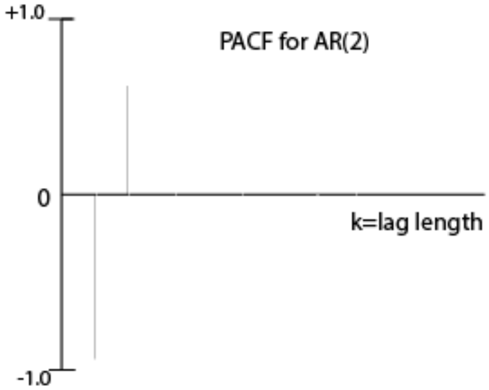
Example I:



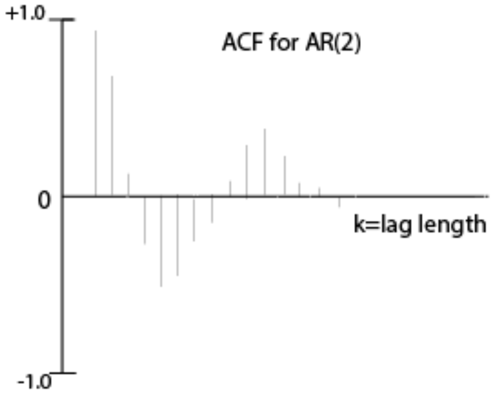


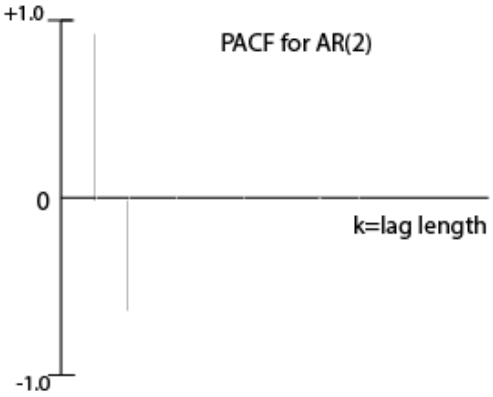
Example II:





Example III:





Example IV:

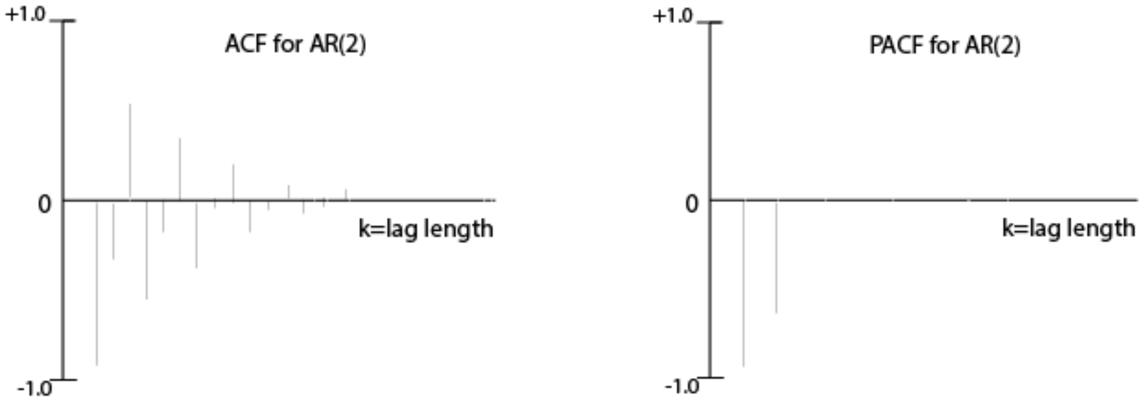


Figure Seven: Theoretical ACF and PACF's for AR (2) model

Stationarity

Earlier it was mentioned that we require our models to be stationary. In order to have stationarity, there are conditions that AR coefficients must satisfy. These conditions, and how to determine if a realization or model is stationary in practice are discussed below. (Pankratz , page 130)

The table displayed below explains the conditions of stationarity for AR coefficients: (Pankratz, page 130)

Model Type	**Stationarity Conditions**
ARMA (0, *q*)	Always stationary

AR (1) or ARMA (1, *q*)	$\lvert\square_1\rvert < 1$
AR (2) or ARMA (2, *q*)	$\lvert\square_2\rvert < 1;\ \square_2 + \square_1 < 1$ $\square_2 - \square_1 < 1$

For *p* >2, the stationarity conditions become complicated. However, this doesn't happen frequently in practice, so we omit discussion on the complicated situation.

Ensuring the stationarity of a series can help us generate useful estimates of the mean, variance and acf. Since there is only one observation at time *t*, it would be difficult to produce a practical model if the mean were changing over time. Therefore, it is important to have the following check for the stationarity. First, examine the realization visually to see if the mean and variance appear to be constant, and examine the estimated acf to see if it drops to zero rapidly; if it does not, the mean may not be stationary and differencing maybe needed. Finally, check any estimated AR coefficients to see that they meet the relevant stationarity conditions. (Pankratz, page 149)

Invertibility

Like AR process, there are certain conditions on MA coefficient that must be satisfied. The table below gives a description of these certain conditions.

Model Type	Invertibility Conditions
ARMA (*p*, 0)	Always invertible

MA (1) or ARMA (*p*, 1)	$\lvert\theta_1\rvert < 1$
MA (2) or ARMA(*p*, 2)	$\lvert\theta_2\rvert < 1; \theta_2 + \theta_1 < 1$ $\theta_2 - \theta_1 < 1$

It is significant to have $|\theta| < 1$, since this condition can lead to invertibility. Recall equation (12):

$$\text{MA (1): } \tilde{z}_t = (1 - \theta_1 B)a_t$$

If $|\theta| < 1$, we can write the MA (1) as the following:

$$(1 - \theta_1 B)^{-1}\tilde{z}_t = a_t$$

and further, write it into a series:

$$\sum_{j=0}^{\infty}(\theta_1)^j \cdot z_{t-j} = a_t$$

which is invertible according to the definition.

The definition of invertibility stated explicitly that a linear process $\{z_t\}$ is invertible (strictly an invertible function of $\{a_t\}$ if there is a

$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \cdots$$

with $\sum_{j=0}^{\infty}\left|\pi_j\right| < \infty$ and $a_t = \pi(B)z_t$ (Bartlett).

The implement of invertibility can help us get more accurate weights placed on past *z* observations. Observations further the past have less effect on the present. Therefore, if the coefficients for the equivalent AR process of infinitely high order do not decline, then the invertibility of this series is violated.

CHAPTER FOUR

Estimation Stage

Principle of Estimation

In the identification stage, we estimate *n/4* autocorrelation and partial autocorrelation coefficients, and then tentatively choose a model. For the estimation stage, we precisely estimate a small number of parameters of the chosen model. At the estimation stage, the values of estimated coefficients are chosen based on some criteria. Maximum likelihood (ML) and least-squares (LS) are two favored criteria chosen by Box and Jenkins. Another common methods used is a *nonlinear least-squares* (NLS) approach. We will introduce these methods used to estimate the coefficients in the next discussion.

"Least-squares" refers to parameter estimates associated with the smallest sum of squared residuals or errors. A residual is the difference between the true value and the predicted one. For example, for an AR (1) model:

$$\left(1-\square_1 B\right)\tilde{z}_t = a_t$$

or

$$z_t = \square(1-\square_1) + \square_1 z_{t-1} + a_t \qquad (15)$$

where $\mu(1-\phi_1)$ is the constant term C. If we want to predict the value of z_t, we need μ, ϕ_1 and z_{t-1}. Since a_t is the random shock, it is not observable. Thus, we assign the expected value of a_t to be zero and use μ, ϕ_1, and z_{t-1} to find the predicted value of z_t designated $\hat{z}_t$:

$$\hat{z}_t = \mu(1-\phi_1) + \phi_1 z_{t-1} \qquad (16)$$

Then we subtract equation (16) from (15), and get:

$$z_t - \hat{z}_t = a_t$$

In practice, we do not know the true values for our parameters of ARIMA models; but we use the estimated ones from the data, i.e. $\hat{\mu}$ and $\hat{\phi}_1$. Therefore, the modified equation (16) is like the following:

$$\hat{z}_t = \hat{\mu}(1-\hat{\phi}_1) + \hat{\phi}_1 z_{t-1} \qquad (17)$$

The residual is the estimated random shock $\hat{a}_t$, which is the difference between equation (17) and (15).

Next, we introduce the concept SSR---sum of squared residuals. First, let us take a look at how to calculate SSR.

The table below is an example of an AR (1), with $\hat{\phi}_1 = 0.5$.

t	z_t	$\tilde{z}_t = z_t - \bar{z}$	$\hat{\tilde{z}}_t = \hat{\phi}_1 \tilde{z}_{t-1}$	$\hat{a}_t = \tilde{z}_t - \hat{\tilde{z}}_t$	$\hat{a}_t^2$

0	/	/	/	/	/
1	80	20	/	/	/
2	60	0	10	-10	100
3	30	-30	0	-30	900
4	40	-20	-15	-5	25
5	70	10	-10	20	400
6	80	20	5	15	225

From column 2, we can calculate $\widehat{\square} = \bar{z} = \frac{\sum z_t}{n} = 60$. Recall that $\tilde{z}_t$ is the differenced series with the same stochastic properties as z_t, but a mean at zero, i.e. $\sum \tilde{z}_t = 0$. The last column showed the squared residuals, moreover, we can calculate SSR from it: $SSR = \sum \hat{a}_t^2 = 1650$.

Our task is to assign different values on $\widehat{\square}_1$ that will give us the minimum SSR.

Nonlinear Least-squares Estimation

Unlike linear-least-squares (LLS) methods, where we derive a set fairly easy-solving of equations from SSR using calculus, with an ARIMA model SSR produces a set of highly nonlinear equations and can be solved only with a nonlinear, iterative technique. This technique is similar to the “*grid-search*” method.

The grid search is idea from of trial-and-error. From the example above, using an AR (1) model, we got a SSR of 1650 by assigning $\widehat{\square}_1 = 0.5$. We can also try other different values from -1 to1, since $|\widehat{\square}_1| < 1$ is promised under a stationary condition. Therefore, we can plot each corresponding SSR associates with the change of $\widehat{\square}_1$, shown as below:

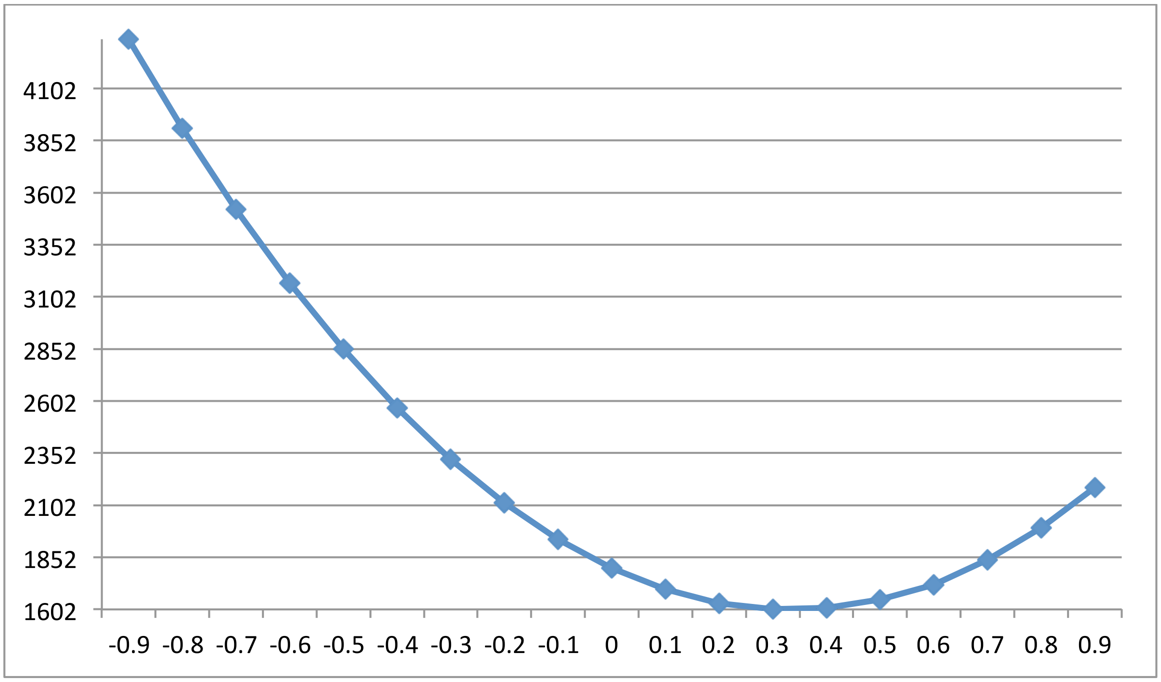


Figure Eight: Nonlinear least squares estimation-grid search

Obtained from this graph, we now have a sense that the minimum SSR appears between $\widehat{\square}_1 = 0.2{\sim}0.5$. Then, we can assign a set of different values from this range to $\widehat{\square}_1$, calculate SSR and compare and so forth. The commonly used NLS method, however, has a better algorithm and will converge quicker to the ideal the estimates using a computer.

Maximum Likelihood Estimation of ARIMA Models

This approach is a common statistical approach to estimate. However, in time series we cannot use the normal version of the maximum likelihood estimation (MLE), but a modified one. First, we will review the normal version and introduce the new one for an ARIMA model.

To find the maximum likelihood function for an identically independent data set (*iid*) with marginal probability density function (pdf) $f(y_t;\theta)$, we first find the joint pdf, which is just the product of the marginal pdf:

$$f(y;\theta) = f(y_1, \dots, y_T;\theta) = \prod_{t=1}^{T} f(y_t;\theta) \qquad (18)$$

The likelihood function is this joint density but treating θ as the parameter given the data y:

$$L(\theta|y) = L(\theta|y_1, \dots, y_T) = \prod_{t=1}^{T} f(y_t;\theta) \qquad (19)$$

Then we take the natural log of both sides and get the log-likelihood:

$$\ln L(\theta|y) = \sum_{t=1}^{T} \ln f(y_t; \theta) \qquad (20)$$

However, for a time series data set, this method does not work because the random variables in the sample are not *iid*. One approach is to use the T×T variance covariance matrix $\mathrm{var}(\mathrm{y})$ to determine the joint pdf $f(y_1, \ldots, y_T; \theta)$ directly. Discussed below is an alternative solution to this problem. We can write the joint density as a product of conditional densities and then maximize the log likelihood (Zivot). Starting from only two observations of a time series y_1, y_2, the conditional density of y_2 given y_1 we obtain

$$f(y_1, y_2; \theta) = f(y_2|y_1; \theta) f(y_1; \theta) \qquad (21)$$

Carrying the same pattern, for three observations, the conditional density becomes:

$$f(y_1, y_2, y_3; \theta) = f(y_3|y_2, y_1; \theta) f(y_2|y_1; \theta) f(y_1; \theta) \qquad (22)$$

In general, for T observations, this conditional density form becomes like this:

$$f(y_1, \ldots, y_T; \theta) = \left(\prod_{t=p+1}^{T} f(y_t|I_{t-1}, \theta) \right) \cdot f(y_p, \ldots, y_1; \theta) \qquad (23)$$

where $I_t = \{y_t, \ldots, y_1\}$ denotes the information available at time *t*, and $y_p, \ldots, y_1$ denotes the initial values. Then the log-likelihood can be written as the following:

$$\ln L(\theta|y) = \sum_{t=p+1}^{T} \ln f(y_t|I_{t-1}, \theta) + \ln f\big(y_p, \ldots, y_1; \theta\big) \qquad (24)$$

This is called the exact log-likelihood. The first term with summation is called the conditional log-likelihood, and the second term is called the marginal log-likelihood for the initial values. Therefore, there are two types of maximum likelihood estimates (mles) that can be computed from equation (24). The first is based on the first term and the estimators are called conditional mles given by

$$\hat{\theta}_{cmle} = \max_{\theta} \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \theta) \qquad (25)$$

The second type is based on the exact log-likelihood function. We call this estimator exact mles. Here is the definition:

$$\hat{\theta}_{mle} = \max_{\theta} \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \theta) + \ln f(y_p, \dots, y_1; \theta) \qquad (26)$$

It is important for us to keep in mind that these two types have the same limiting normal distribution and the two estimators $\hat{\theta}_{cmle}$ and $\hat{\theta}_{mle}$ are consistent if we have stationary models. It should be noted that "$\hat{\theta}_{cmle}$ and $\hat{\theta}_{mle}$ can be substantially different if the series is non-stationary or non-invertible." (Zivot)

We now demonstrate the maximum likelihood estimation in the following example. Consider an AR (1) model with the assumption of the random shock term a_t~iid $\mathrm{N}(0, \square^2)$:

$$y_t = c + \square_1 y_{t-1} + a_t, a_t \sim \text{iid } \mathrm{N}\left(0, \square^2\right), t = 1, \dots, T \qquad (27)$$

$$\boldsymbol{\theta} = \left[c, \phi, \sigma^2\right]', |\phi| < 1$$

Then conditional on I_{t-1} is

$$y_t | I_{t-1} \sim N\left(c + \phi y_{t-1}, \sigma^2\right), t = 2, \dots, T$$

which only depends on y_{t-1}. The conditional density $f(y_t | y_{t-1}, \boldsymbol{\theta})$ is then

$$f(y_t | y_{t-1}, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_t - c - \phi y_{t-1})^2\right), t = 2, \dots, T \qquad (28)$$

Next, we determine the marginal density for the initial value y_1. In order to do that, we first determine the mean and variance of the AR (1) process. Recall that $|\phi|$ is less than 1 under stationarity or the variance goes to infinity. Consequently, the mean $\mathrm{E}(y_t)$ is identical for all t and we denote the mean to be μ. Therefore,

$$E(y_t) = E(c) + \phi_1 E(y_{t-1}) + E(a_t)$$

or

$$\mu = c + \phi_1 \mu$$

As a result,

$$\mathrm{E}(y_1) = \mu = \frac{c}{1 - \phi} \qquad (29)$$

Also, we can determine the variance, which is

$$\mathrm{Var}(y_1) = \frac{\sigma^2}{1-\phi^2} \qquad (30)$$

Then $y_1 \sim N\left(\frac{c}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right)$ and its density is

$$f(y_1;\, \boldsymbol{\theta}) = \left(2\pi\frac{\sigma^2}{1-\phi^2}\right)^{-1/2} exp\left(-\frac{1-\phi^2}{2\sigma^2}(y_1 - \frac{c}{1-\phi})^2\right) \qquad (31)$$

Thus,

$$f(y_1, \dots, y_T; \theta) = \left(\prod_{t=2}^{T} f(y_t | I_{t-1}, \theta)\right) \cdot f(y_1; \theta)$$

$$= (2\pi\sigma^2)^{-(T-1)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2\right) \cdot f(y_1; \theta)$$

Then the conditional log-likelihood function is

$$\sum_{t=2}^{T} \ln f(y_t | y_{t-1}, \boldsymbol{\theta}) = -\frac{T-1}{2}\ln(2\pi)$$

$$-\frac{T-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 \qquad (32)$$

If we take the derivative of equation (32) with respect to σ^2 and set it to zero, then the solution is our conditional mle for σ^2:

$$\widehat{\sigma}^2_{cmle} = (T-1)^{-1}\sum_{t=2}^{T}(y_t - \hat{c}_{cmle} - \widehat{\phi}_{cmle} y_{t-1})^2 \qquad (33)$$

Then we take the log function of the both sides of equation (31) and get the marginal log-likelihood for the initial value y_1 is

$$\ln f(y_1;\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\square) - \frac{1}{2}\ln\left(\frac{\square^2}{1-\square^2}\right) - \frac{1-\square^2}{2\square^2}\left(y_1 - \frac{c}{1-\square}\right)^2 \qquad (34)$$

Further, the exact log-likelihood function is the addition of equation (32) and (34):

$$\ln L(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{T}{2}\ln(2\square) - \frac{1}{2}\ln\left(\frac{\square^2}{1-\square^2}\right) - \frac{1-\square^2}{2\square^2}\left(y_1 - \frac{c}{1-\square}\right)^2$$

$$-\frac{T-1}{2}\ln(\square^2) - \frac{1}{2\square^2}\sum_{t=2}^{T}(y_t - c - \square y_{t-1})^2 \qquad (35)$$

Since the exact log-likelihood function is a non-linear function of the parameter $\boldsymbol{\theta}$, there is no closed form solution for the exact mles. As Zivot pointed out that the exact mles must be determined by numerically maximizing the exact log-likelihood function. Usually we use an iterative scheme from the Newton-Raphson type algorithm for the maximation:

$$\widehat{\boldsymbol{\theta}}_{mle,n} = \widehat{\boldsymbol{\theta}}_{mle,n} - \widehat{\boldsymbol{H}}\left(\widehat{\boldsymbol{\theta}}_{mle,n-1}\right)^{-1}\widehat{\boldsymbol{s}}\left(\widehat{\boldsymbol{\theta}}_{mle,n-1}\right) \qquad (36)$$

"where $\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{\theta}})$ is an estimate of the Hessian matrix (second derivative of the log-likelihood function), and $\widehat{\boldsymbol{s}}(\widehat{\boldsymbol{\theta}})$ is an estimate of the score vector (first derivative of the log-likelihood function). The estimates of the Hessian and Score may be computed numerically (using numerical derivative routines) or they may be computed analytically (if analytical derivatives are known)."

CHAPTER FIVE

Forecasting

To forecast future values for an ARIMA model, we have point estimator and interval estimators. In this chapter, we introduce the method of producing the point estimate for ARIMA forecasts and then we consider interval estimation. Further, we illustrate some criteria for an optimal forecast.

The most convenient way to derive the point estimators are to write the model in difference-equation form. Let t be the current time period, and we are interested in forecasting l periods after t, in which t is also called the forecast origin and l is called the forecast lead time. z_{t+l} is denoted to time period l after the origin. Then the forecast of z_{t+l}, designated as $\hat{z}_t(l)$, is the conditional expectation of z_{t+l}:

$$\hat{z}_t(l) = E(z_{t+l}|I_t) \qquad (37)$$

where I_t is the information contained in the set of pervious observations $(z_t, z_{t-1}, \dots)$. An example of an ARIMA model (1, 0, 1) will illustrate this better. Written in the backshift notation, ARIMA (1, 0, 1) model is

$$(1-\phi_1 B)\tilde{z}_t = (1-\theta_1 B)a_t$$

or

$$z_t = \mu(1-\phi_1) + \phi_1 z_{t-1} - \theta_1 a_{t-1} + a_t$$

Then for time period *t+1* (where *l=1*), the expression is

$$z_{t+1} = \mu(1-\phi_1) + \phi_1 z_t - \theta_1 a_t + a_{t+1} \qquad (38)$$

Then we apply equation (37) to (38) to find the forecast for z_{t+1}:

$$\hat{z}_t(1) = E(z_{t+1}|I_t) = \mu(1-\phi_1) + \phi_1 z_t - \theta_1 a_t \qquad (39)$$

Since we do not know a_{t+1} at time *t*, we assign zero for its expected value. Thus, z_t and a_t are relevant to the information needed for z_{t+1} prediction.

In practice, we use the previous predicted value $\hat{z}_t$ to forecast z for the next time period $\hat{z}_{t+1}$ and forecasts for *l* > *1* are called "bootstrap" forecasts. Similarly, the mean μ, coefficients ϕ and θ are unknown as well, so we use estimated values for them, i.e. $\hat{\mu}, \hat{\phi}$ and $\hat{\theta}$. A forecast for the 61th and 62th period with a sample size n=60 can be written as:

$$\hat{z}_{61} = \hat{z}_{60}(1) = \hat{C} + \hat{\phi}_1 z_{60} - \hat{\theta}_1 \hat{a}_{60} \qquad (40)$$

$$\hat{z}_{62} = \hat{z}_{60}(2) = \hat{C} + \hat{\phi}_1 \hat{z}_{61} - \hat{\theta}_1 \hat{a}_{61} \qquad (41)$$

where $\hat{C} = \hat{\mu}(1-\hat{\phi}_1)$ is the estimated constant term.

If we assign some particular values to $\hat{\mu}, \hat{\phi}$ and $\hat{\theta}$, then our forecasts are converging toward the mean of the series as long as the model is stationary. Whether the convergence is rapid or slow depends on each individual model. When the lead time (*l*) exceeds the lag length of a past shock term (q), we have seen that with a pure AR or mixed models, we use forecast *z*'s instead of observed *z*'s, which we referred to as "bootstrap". However, with a pure MA or mixed models, we must replace random shocks with the expected value of zero. Furthermore, when $l > q$, the forecasts are equal to the estimated mean $\hat{\mu}$ in pure MA models.

In order to derive confidence intervals around our point forecasts, we need to write all AR terms with an infinite series of MA terms, since a pure MA model is already in random-shock form.

Symbol ψ is used to denote the coefficients in a random-shock form, with the subscript *t* corresponding to the time lag of the associated past random shock:

$$z_t = \mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \psi_3 a_{t-3} + \cdots \qquad (42)$$

where $\psi_0 = 1$. If this sequence is finite, then it represents a pure MA model. On the other hand, if it is infinite, it represents a pure AR or mixed model. And μ is the mean for a stationary series. For example, an MA (2) can be written as:

$$(z_t - \mu) = (1 - \theta_1 B - \theta_2 B^2) a_t$$

or

$$z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

If we let $\psi_0 = 1,\ \psi_1 = -\theta_1$ and $\psi_2 = -\theta_2$, we can rewrite the above equation in random-shock form as

$$z_t = \mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2}$$

which is a truncated version of equation (42). Therefore, the random-shock form will make it easier for our later discussion about confidence interval.

Equation (40) and (41) gave us an idea about how to conduct a point estimate process. Now we are interested in how reliable a point estimate is, and we incorporate the variability by constructing a confidence interval. First, we consider the forecast-error, which is the difference between the observed data and the forecasted one, denoted by $e_t(l)$:

$$e_t(l) = z_{t+l} - \hat{z}_t(l) \qquad (43)$$

Then we write the observed *z* for period *t+l* in random-shock form:

$$z_{t+l} = \mu + \psi_0 a_{t+l} + \psi_1 a_{t-1+l} + \psi_2 a_{t-2+l} + \psi_3 a_{t-3+l} + \cdots \qquad (44)$$

and the corresponding forecast term $\hat{z}_t(l)$ can be written as:

$$\hat{z}_t(l) = E(z_{t+l}|I_t) = \square + \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots \qquad (45)$$

since the random shock terms a_t's cannot be known beyond time *t*. Therefore, when we substitute equation (44) and (45) into equation (43) to get

$$e_t(l) = \psi_0 a_{t+l} + \psi_1 a_{t-1+l} + \psi_2 a_{t-2+l} + \cdots + \psi_{l-1} a_{t+1} \qquad (46)$$

Next, we can use equation (46) to find the conditional variance of $e_t(l)$ is

$$\square^2[e_t(l)] = E\{e_t(l) - E[e_t(l)]|I_t\}^2 = E[e_t(l)]^2$$

$$= \sigma_a^2(1 + \psi_1{}^2 + \psi_2{}^2 + \cdots + \psi_{l-1}{}^2)$$

As showed earlier, $E(a_{t+1}) = E(a_{t+2}) = \cdots = 0$. In addition, all cross-product terms have an expected value of zero since the random-shocks are assumed to be independent. Constant σ_a^2 is the expected value for each of the remaining squared shock terms according to our assumption.

Further, the standard deviation of $e_t(l)$ is the following:

$$\sigma[e_t(l)] = \sigma_a(1 + \psi_1{}^2 + \psi_2{}^2 + \cdots + \psi_{l-1}{}^2)^{1/2}$$

However, this result is not exact. The standard deviation can only be approximated since σ_a^2, ψ_i are estimated.

Having calculated the standard deviation, we can then construct confidence intervals. We made two assumptions here: one, our sample size is large enough and two, the random shocks are normally distributed in order to apply the central limit theorem---our forecasts from the model are approximately normally distributed. Thus, an approximate 95% confidence interval is given by

$$\hat{z}_t(l) \pm 1.96\widehat{\sigma}[e_t(l)]$$

More generally, an approximate 100(1- α) % confidence interval is:

$$\hat{z}_t(l) \pm Z_{\frac{1-\alpha}{2}} \cdot \widehat{\sigma}[e_t(l)]$$

where $Z_{\frac{1-\alpha}{2}}$ is the lower percentile of a standard normal distribution.

CONCLUSION

This thesis takes a brief look at the modeling of time series data using the Univariate Box-Jenkins model, and introduces readers to its three-stage process. We did not consider diagnostic checking, which is often involved in practice before the forecasting stage to make sure a model is legitimate and optimal. Also, our discussion on UBJ model is limited to only stationary data set. More processes will be needed if the data set is non-stationary, like differencing. Further work could include comparing different methods of estimation. Bayesian methods are of particular interest.

REFERENCE

Bartlett, Peter. "Introduction to Time Series Analysis. Lecture 5." Berkeley. 20 Apr. 2014. University of California, Berkeley. Web. 20 Apr. 2014.

Brillinger, David R. "Time Series: General." (2000): n. pag. Rpt. in Int. Encyc. Social and Behavioral Sciences. Berkeley: U of California, 2000. Print.

Easton, Valerie J., and John H. McColl. "Statistics Glossary - Time Series Data." *Statistics Glossary - Time Series Data*. Stuart G. Young, n.d. Web. 14 Apr. 2014.

"Estimation of ARMA Models." UW Faculty Web Server. University of Washington, 6 Apr. 2005. Web. 20 Mar. 2014.

G. E. P. Box (1983) G. M. Jenkins, 1933-1982 Journal of the Royal Statistical Society. Series A (General), Vol. 146, No. 2, pp. 205-206.

Pankratz, Alan. *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. New York: Wiley, 1983. Print.

"Statistics Toolbox." - *Time Series Regression of Airline Passenger Data Demo*. N.p., n.d. Web. 17 Apr. 2014.

Wasserstein, R. (2010), “George Box: A Model Statistician”, Significace, 7(3), 134-135.

Zivot, Eric. "Estimation of ARMA Models." 6 Apr. 2005. University of Washington. Web. 1 Mar. 2014.