Time Series

M2R Project

Amir Rahman, Jonathan Rubin, Joshua Tegegne
Selena Linden, Walter Shen

Supervisor: Henrique Helfer Hoeltgebaum
Group 45

Department of Mathematics
Imperial College London
June 2021

Contents

1 Introduction

1.1 What is a time series?

A times series is a collection of data points $\{y_t\}_{t=1}^{T}$ which track the value of a particular quantity over time. Usually these measurements are taken in equally spaced intervals, but this is not necessarily required. For the purposes of this report, we will only deal with cases where the spacing between measurements is equal.

Examples of uses of time series are in finance, where we can model the returns on a particular stock, or in epidemiology, where we can model the number of new daily cases of an infection, such as COVID-19. Time series can be used to forecast the value of the quantity in question at future times, however there are of course limitations, some of which will be discussed over the course of this report.

1.2 IBM Stock Data

Throughout the report, we will often make reference to "real world data sets", and to help visualise how the analysis proposed would be carried out, we have chosen to analyse the IBM stock data. In particular, we will be looking at the lowest stock price each day in the time period 2006-2017. [1]

Due to the nature of this data, we will introduce the method of Differencing in section 9, which we will apply to the original data. Therefore we will refer to both the original "IBM Time Series" and the "Once-differenced IBM Time Series".

1.3 What is an Autoregressive Time Series Model?

There are several ways in which the value of the time series at time T can be estimated. In an autoregressive time series, we take the value of the time series at time T to be the sum of a constant term, ϕ_0, a linear combination of a select number of previous terms, and a Gaussian error term, ε_T. The constant term and error are present in all autoregressive time series, but the number of past terms used is dependent on the order. A few examples and properties of autoregressive time series are presented in sections 1.3.1 and 1.3.2.

1.3.1 AR(1) and AR(2) Time Series

In an AR(1) time series, the present value is dependent only on the previous value. Explicitly written, it can be said that

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

Taking this one step further, we can define the autoregressive model of order 2 by considering the next step backwards in the following way:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

For both of these models, the properties are very easily obtained, and throughout this report, we will regularly refer to these two models as starting points for obtaining more general results.

1.3.2 AR(p) Time Series

An autoregressive time series doesn't have to be dependent on just the first and second terms backwards in time. In particular, for monthly data which shows seasonal trends each year, it is logical to believe that the present value is dependent on the value 12 time periods ago, otherwise seen as the same time in the previous year.

This is not the only example of cases where further back terms are used, and so this motivates us to define a more general model, the AR(p) model, where we evaluate the present time value as a function of the previous p values as follows:

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma^2)$$

For all of these models, the expectation and variance of the time series are crucial for analysing the data, and deciding whether a time series is suitable for the data presented.

It is important to note at this point that sometimes we will include the term ϕ_0, and sometimes we won't. If we ever do want to remove this constant from the model to make it easier to work with in a linear sense, then we can define a new time series $\tilde{y}_t = y_t - \mu$ where $\mu = \mathrm{E}(y_t)$. This centres the time series around 0 and from this, we can immediately estimate $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$

1.3.3 Errors

The errors are included in this model to represent the fact that in real world data, it is practically impossible to have a set of data that is exactly a linear combination of the p previous values. One example of where this can be observed is at the start of the 1991 Gulf War. The war was an extenuating external factor that caused a large, though temporary, spike in the price of crude oil, due to a drop in production rate [2].

The errors in the AR(p) model are used to mimic these real world shocks and allow for a smoother linear component. Here, the initial shock created by the error does not have a time limited impact, however its impact can be seen to reduce over time. The errors are independent and identically distributed with a normal distribution with mean 0 and variance σ_ε^2. This feeds into defining and proving some of the key features of estimating the parameters and the correlation functions.

2 Stationarity

2.1 Strictly Stationary

A time series is said to be strictly stationary if the joint distribution of $(y_1, ..., y_t)$ is invariant under time shifts. More specifically, $(y_1, ..., y_t)$ has to be identically distributed as $(y_{1+\tau}, ..., y_{t+\tau})$ for all τ, where t is an arbitrary positive integer [3, p. 25]. This condition is usually very hard to impose and check, so we revert a less restrictive form of stationarity known as *Weak Stationarity*.

2.2 Weakly Stationary

A time series y_t is weakly stationary if both the mean of y_t and the covariance between y_t and $y_{t-\ell}$ remain constant for all t , where ℓ is an arbitrary integer [3, 25]. For this, we require for all t:

- $\mathrm{E}(y_t) = \mu$
- $\mathrm{Cov}(y_t, y_{t-\ell}) = \gamma_\ell$

Here there is an underlining assumption that the first and second moments of y_t exist. Under this assumption, strict stationarity implies weak stationarity. The converse is not true however [3, p. 25].

The covariance $\gamma_\ell = \mathrm{Cov}(y_t, y_{t-\ell})$ is called the lag-ℓ autocovariance of y_t. Two simple but important properties of the autocovariance are:

- $\gamma_0 = \mathrm{Var}(y_t)$
- $\gamma_\ell = \gamma_{-\ell}$

These are true because firstly, $\gamma_0 = \text{Cov}(y_t, y_t) = \text{Var}(y_t)$ and secondly $\gamma_{-\ell} = \text{Cov}(y_t, y_{t+\ell}) = \text{Cov}(y_{t-\ell}, y_t) = \gamma_\ell$ (We have used the weak stationarity assumption).

2.2.1 Mean and Variance of Stationary AR Models

Under the conditions of weak stationarity, we can calculate explicit formulae for the Mean and variance of AR models [4].

- $\text{E}(y_t) = \frac{\phi_0}{1-\phi_1-\phi_2-\ldots-\phi_p}$
- $\text{Var}(y_t) = \sigma^2 + \sum_{i=1}^{p} \phi_i \gamma_i$

3 White Noise

One of the first things we look for when analysing a potential time series is answering the fundamental question "Can we fit a time series to this data?" This puts aside any question of best modelling approach, and focuses on the core properties of the data we are given. For a given set of data $\{y_t\}_{t=1}^{T}$, the autocovariance at lag ℓ, γ_ℓ, is defined by $\gamma_\ell = \text{E}[(y_t - \mu)(y_{t+\ell} - \mu)]$.

Now if we consider a set of independent, identically distributed random variables, $\{\varepsilon_t\}_{t=1}^{T}$, all with zero mean and fixed variance σ^2, then using the fact that the ε_t are independent, and therefore uncorrelated, we have for the autocovariance function that

$$\gamma_\ell = \text{E}[\varepsilon_t \varepsilon_{t+\ell}] = \begin{cases} \sigma^2 & \ell = 0 \\ 0 & \ell \neq 0 \end{cases}$$

If the data satisfies the criteria and autocovariance function detailed above, or is similar to it, then we call this *white noise*. If a set of data is seen to be white noise, then it is not possible to fit a time series to the data, as there is not enough evidence to suggest that the y_t are correlated, and so a time series of any sort would not be suitable. Below shows how a white noise time series looks on a histogram and a QQ Norm plot. The shapes of these plots will be used further in the report to analyse residuals.

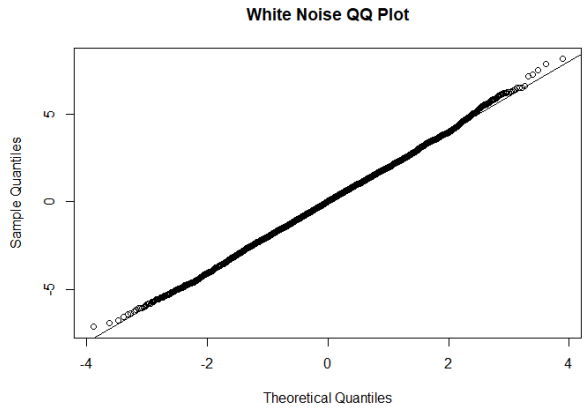

Figure 1: Normal QQ Plot of White Noise

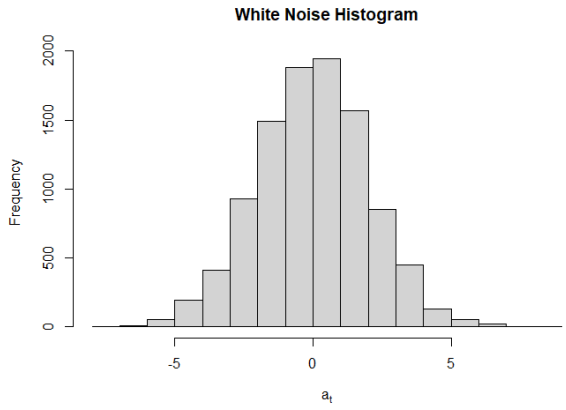

Figure 2: Histogram of White Noise

4 Autocorrelation Function (ACF)

4.1 Motivation and Definition

A further way to answer the question "Can we fit a time series to this data?" is to analyse the *autocorrelation function* (ACF). Indeed, the ACF for white noise and autoregressive time series exhibits particular characteristics. Therefore, analysing the ACF gives us insight into which model is most appropriate for a given time series.

The ACF captures the linear dependence between y_t and past values $y_{t-\ell}$ in a time series $\{y_t\}_{t=1}^T$. Mathematically, the ACF is defined as the collection $\{\rho_\ell, \ell \in \mathbb{N}\}$, where ρ_ℓ is the correlation coefficient between y_t and $y_{t-\ell}$, also known as the *lag-ℓ autocorrelation* of y_t [5, p. 2]. Thus, for a weakly stationary time series, the lag-ℓ autocorrelation is closely related to the autocovariances at lag ℓ and 0, defined in section 2.2:

$$\rho_\ell = \frac{\mathrm{Cov}(y_t, y_{t-\ell})}{\sqrt{\mathrm{Var}(y_t)\mathrm{Var}(y_{t-\ell})}} = \frac{\mathrm{Cov}(y_t, y_{t-\ell})}{\mathrm{Var}(y_t)} = \frac{\gamma_\ell}{\gamma_0}$$

as $\mathrm{Var}(y_t) = \mathrm{Var}(y_{t-\ell})$ under the weak stationarity assumption [3, p. 26].

In practice, when given a time series sample $\{s_t\}_{t=1}^T$ with sample mean $\bar{s}$, one can look instead at the *lag-ℓ sample autocorrelation* of s_t, defined as [3, p. 26] :

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^T (s_t - \bar{s})(s_{t-\ell} - \bar{s})}{\sum_{t=1}^T (s_t - \bar{s})^2}$$

$\hat{\rho}_\ell$ is a biased estimator of ρ_ℓ, with bias of order $\frac{1}{T}$ [3, p. 27]. Therefore in large finite samples it is reasonable to use. To visualise the behaviour of the ACF of sample time series $\{s_t\}_{t=1}^{T}$, one can then use an *autocorrelogram*, which is a plot of the lag-ℓ sample autocorrelations against lag ℓ.

4.2 ACF of white noise time series

It follows that the lag-ℓ autocorrelations are all zero for a white noise time series, as it is a sequence of independent random variables.

In practice, plotting an autocorrelogram of a time series and observing that the lag-ℓ sample autocorrelations are not significantly different from zero suggests a white noise time series, and therefore an incompatibility with a time series model. Note that, as an exception, for lag 0, we have that $\rho_0 = \frac{\gamma_0}{\gamma_0} = \frac{\sigma^2}{\sigma^2} = 1$.

By convention, the sample lag-ℓ autocorrelations are considered not significantly different from zero if they lie in the two-standard error interval $\left(-\frac{2}{\sqrt{T}}, \frac{2}{\sqrt{T}}\right)$. Indeed, the lag-$\ell$ sample autocorrelation follows asymptotically a normal distribution with mean equal to its population mean and standard deviation $\frac{1}{\sqrt{T}}$ [5, p. 8]. At a 5% significance level, the standard normal 2-tailed test statistic comes out to be roughly equal to 1.96. Hence, rounding up 1.96 to 2, the sample lag-ℓ autocorrelations are considered negligible if the values fall in the 95% confidence interval $\left(-\frac{2}{\sqrt{T}}, \frac{2}{\sqrt{T}}\right)$ centred at 0.

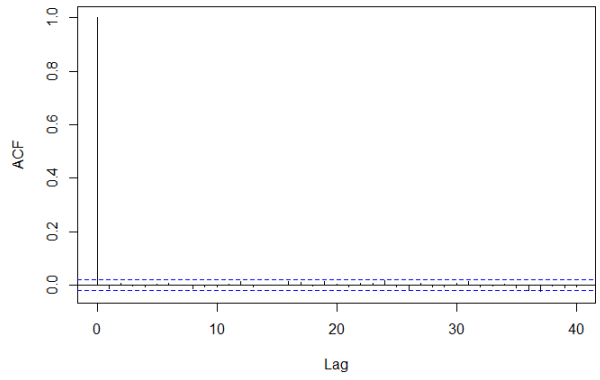

Figure 3: ACF of White Noise

4.3 ACF of autoregressive time series

For weakly stationary autoregressive models, the ACF takes on notable forms.

4.3.1 AR(1) time series

For the stationary AR(1) model $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$, with $\text{Var}(\varepsilon_t) = \sigma^2$, we have that [3, p. 34]:

$$\gamma_\ell = \begin{cases} \dfrac{\sigma^2}{1-\phi_1^2} & \ell = 0 \\ \phi_1 \gamma_{\ell-1} & \ell > 0, \end{cases}$$

which is known as the *moment equation* of a stationary AR(1) model. Upon dividing by γ_0, we obtain a recursive relation for the ACF. From this relation, we can deduce a general solution for ρ_ℓ :

$$\rho_\ell = \phi_1 \rho_{\ell-1} \quad \text{for} \quad \ell > 0, \quad \rho_0 = 1 \implies \rho_\ell = \phi_1^\ell \text{ for } \ell \geq 0$$

Therefore, for an AR(1) time series, an autocorrelogram would show an exponential decay for positive rate ϕ_1 and an alternating exponential decay for negative rate ϕ_1, as shown in Figure 4 and Figure 5 respectively.

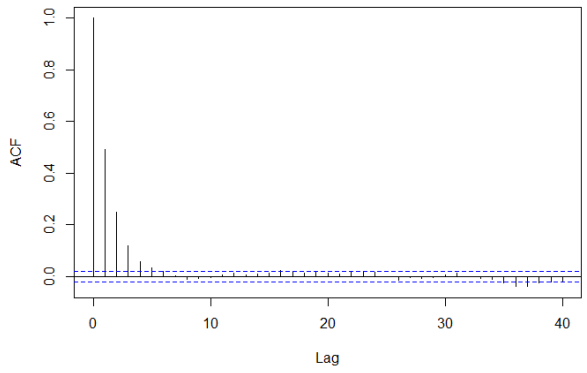

Figure 4: ACF of Time Series with Positive Decay

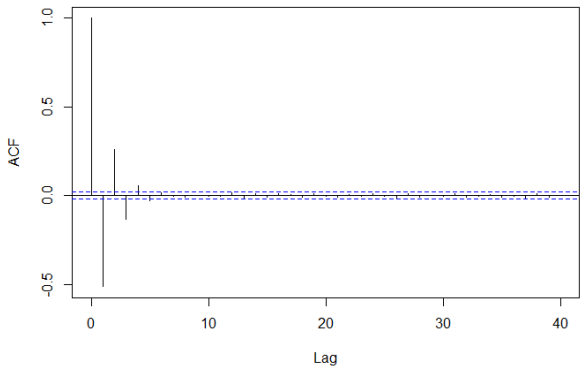

Figure 5: ACF of Time Series with Negative Decay

4.3.2 AR(2) time series

For the AR(2) stationary model $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$, the moment equation is $\gamma_\ell = \phi_1 \gamma_{\ell-1} + \phi_2 \gamma_{\ell-2}$ and the corresponding ACF recursive relation $\rho_\ell = \phi_1 \rho_{\ell-1} + \phi_2 \rho_{\ell-2}$ [3, p. 36].

Using the back-shift operator B, where $B\rho_\ell = \rho_{\ell-1}$, this leads to a second order difference equation $(1 - \phi_1 B - \phi_2 B^2)\rho_\ell = 0$ [6, p. 30] and a corresponding second-order polynomial equation $1 - \phi_1 x - \phi_2 x^2 = 0$, called the *characteristic polynomial* of the AR(2) model [3, p. 36].

Define $x_{1,2}$ as the solutions of the characteristic polynomial and $\omega_i = \frac{1}{x_i}$ for $i = 1, 2$. Then, the ω_i's are known as the *characteristic roots* of the AR(2) model [3, p. 36]. Solving the quadratic characteristic polynomial gives:

$$
\begin{aligned}
& 1 - \phi_1 x - \phi_2 x^2 = 0 \\
\implies & x_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \\
\implies & \omega_{1,2} = \frac{1}{x_{1,2}} = \frac{-2\phi_2}{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}
\end{aligned}
$$

Hence, the characteristic roots are complex if and only if $\phi_1^2 + 4\phi_2 < 0$.

The nature of the characteristic roots are crucial as they determine the behaviour of the ACF. Indeed, from the second order difference equation, one can find the general solution of ρ_ℓ, which depends on the characteristic roots [7, p. 24]:

$$
\rho_\ell = \begin{cases}
\dfrac{1}{\gamma_0}(A_1 \omega_1^\ell + A_2 \omega_2^\ell) & \text{for distinct real characteristic roots } \omega_1 \text{ and } \omega_2 \\
\dfrac{1}{\gamma_0}(A_1 + A_2 \ell)\omega^\ell & \text{for repeated real characteristic roots } \omega_1 = \omega_2 = \omega \\
\dfrac{1}{\gamma_0}\left(\displaystyle\sum_{k=1}^{2} A_k |\omega_k|^\ell \left[\cos(\theta_k \ell) + i \sin(\theta_k \ell)\right]\right) & \text{for complex characteristic roots } \omega_k = |\omega_k| e^{i\theta_k}, k = 1, 2
\end{cases}
$$

where A_i's are general constants determined by initial conditions.

As the ACF is defined as the collection $\{\rho_\ell, \ell \in \mathbb{N}\}$ (see section 4.1), one can deduce the behaviour of the ACF from the general solution of ρ_ℓ . Hence, for real distinct or repeated characteristic roots, the ACF exhibits an exponential decay. However, for complex characteristic roots, the ACF consists of a decaying damped sinusoidal variation with average length cycle $\kappa = \frac{2\pi}{\cos^{-1}(\phi_1 / 2\sqrt{-\phi_2})}$ [3, p. 37].

As an example, Figure 6 shows the autocorrelogram of an AR(2) time series with parameters $\phi_0 = 0.3, \phi_1 = 0.3, \phi_2 = -0.2$. As $(0.3)^2 + 4(-0.2) < 0$, this has complex characteristics roots and displays damped cosine and sine fluctuations, with average length cycle $\kappa = \frac{2\pi}{\cos^{-1}(0.3/2\sqrt{0.2})} \approx 4.17$.

4.3.3 AR(p) time series

The results in sections 4.3.1 and 4.3.2 can be generalised for higher order autoregressive models. For a general AR(p) stationary model, the difference equation with back-shift operator B is [3, p. 40]:

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)\rho_\ell = 0 \text{ for } \ell > 0$$

with corresponding characteristic polynomial:

$$(1 - \phi_1 x - \phi_2 x^2 - ... - \phi_p x^p) = 0$$

The general solution for ρ_ℓ is then given by [7, p. 26]:

$$\frac{1}{\gamma_0}(A_1\omega_1^\ell + A_2\omega_2^\ell + ... + A_p\omega_p^\ell)$$

where A_i's are general constants determined by initial conditions and the ω_i's are the characteristic roots.

Therefore, as in the first and second order case, the ACF of the AR(p) stationary model is a mixture of exponential decays and damped sine and cosine variations, dictated by the factorisation of the characteristic polynomial and the nature of the characteristic roots [3, p. 40].

Furthemore, another important property is that weak stationarity of an AR(p) model is equivalent to having all characteristic roots less than one in modulus [6, p. 34].

4.3.4 Analysis of IBM ACF

Figure 7 displays the autocorrelogram of the Once-differenced IBM time series. We will show in section 9.4 that the Once-differenced IBM time series can be fitted to an AR(2) model with estimated parameters $\phi_1 = 0.1165, \phi_2 = -0.0398$, which implies complex characteristic roots as $(0.1165)^2 + 4(-0.0398) < 0$. Note however that the Once-differenced IBM time series ACF in Figure 7 does not match exactly the damped sinusoidal behaviour of the simulated AR(2), showing a limitation of fitting AR models to real-world data.

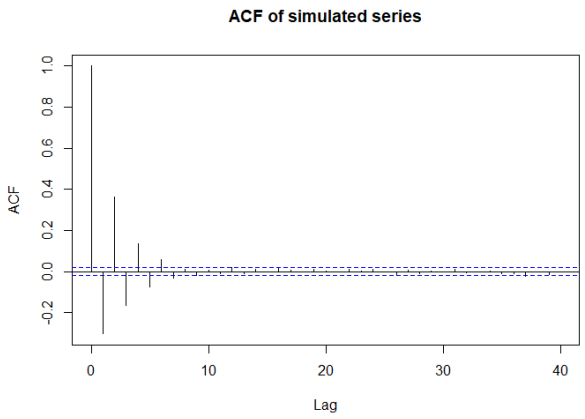

Figure 6: ACF of Simulated AR(2) Time series with Complex Characteristic Roots

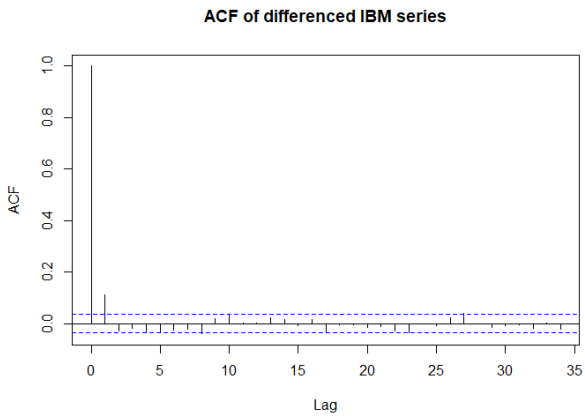

Figure 7: ACF of Once-differenced IBM Time Series

5 Partial Autocorrelation Function (PACF)

5.1 Motivation

While the autocorrelation function is valuable to have, and tells us about linear dependence of y_t on $y_{t-\ell}$ for autocorrelation of lag ℓ, in an autoregressive model, it is more important to consider the correlation between y_t and $y_{t-\ell}$ where we remove dependency on the values of y_{t-1} through to $y_{t-(\ell-1)}$. This is exactly what the partial autocorrelation function can be used for.

Why we define such a function comes directly from the definition of the autoregressive time series. When looking at the modelling the value of y_t, if we consider the value of $y_{t-\ell}$, then we are automatically considering the value of y_{t-i} for $i = 1, \ldots, \ell - 1$, and as part of finding the optimal order of an autoregressive time series, the properties of the time series relate very closely with the PACF.

In simpler terms, the PACF of lag ℓ defines the linear dependence of y_t on $y_{t-\ell}$ which is not accounted for by the terms in-between, namely y_{t-i} for $i = 1, \ldots, \ell - 1$. Viewed more mathematically, the PACF can be written as [8]:

$$\text{PACF}(\ell) = \frac{\text{Cov}(\left[y_t | y_{t-1}, \ldots, y_{t-(\ell-1)}\right], \left[y_{t-\ell} | y_{t-1}, \ldots, y_{t-(\ell-1)}\right])}{\sigma_{\left[y_t | y_{t-1}, \ldots, y_{t-(\ell-1)}\right]} \sigma_{\left[y_{t-\ell} | y_{t-1}, \ldots, y_{t-(\ell-1)}\right]}}$$

The concept of having the term $\left[y_{t-\ell} | y_{t-1}, \ldots, y_{t-(\ell-1)}\right]$ can seem counter-intuitive. Remember that in any time series model, some of the information from time $t-1$ carries through into the value at time t, and through this, y_t can be used as some sort of predictor of y_{t-1}. This logic can be used consecutively to explain the presence of the term $\left[y_{t-\ell} | y_{t-1}, \ldots, y_{t-(\ell-1)}\right]$.

5.2 Calculation

To do the raw calculation of the PACF takes a very long time, and can also become a recursive process, so we will focus on calculating the *sample PACF*. The PACF of lag ℓ looks at the dependence of y_t on $y_{t-\ell}$ given the terms in-between, so a good way to start finding the PACF is to first fit a model of the appropriate order to the data. So to find the lag-ℓ PACF, fit an order-ℓ autoregressive model to the data. So we have

$$y_t = \phi_0 + \sum_{i=1}^{\ell} \phi_i y_{t-i} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

In this model, we need to estimate the quantities $\phi_0, \phi_1, \ldots, \phi_\ell, \sigma^2$. This can be done through any of the methods mentioned in the Parameter Estimation section. Once we have this estimate, we can repeat this for various orders. Typically this is done for orders up to $\frac{T}{4}$,where T is the total number of observations [9, p. 7]. Say we compute p models, and we denote $\phi_{i,j}$ to be the i^{th} parameter for the model of order j, $\phi_{0,j}$ the constant for the model of order j, and σ_j^2 the variance of the errors. We get parameter estimates for the equations:

$$\begin{aligned}
y_t &= \phi_{0,1} + \phi_{1,1} y_{t-1} + \varepsilon_{t,1} && \varepsilon_{t,1} \sim N(0, \sigma_1^2) \\
y_t &= \phi_{0,2} + \phi_{1,2} y_{t-1} + \phi_{2,2} y_{t-2} + \varepsilon_{t,2} && \varepsilon_{t,2} \sim N(0, \sigma_2^2) \\
y_t &= \phi_{0,3} + \phi_{1,3} y_{t-1} + \phi_{2,3} y_{t-2} + \phi_{3,3} y_{t-3} + \varepsilon_{t,3} && \varepsilon_{t,3} \sim N(0, \sigma_3^2) \\
&\vdots \quad \vdots \\
y_t &= \phi_{0,p} + \phi_{1,p} y_{t-1} + \phi_{2,p} y_{t-2} + \phi_{3,p} y_{t-3} + \cdots + \phi_{p,p} y_{t-p} + \varepsilon_{t,p} && \varepsilon_{t,p} \sim N(0, \sigma_p^2)
\end{aligned}$$

From here, we can directly read off the *sample lag-ℓ PACF* to be $\hat{\phi}_{\ell,\ell}$. To see this, start from the order 1 model. $\phi_{1,1}$ describes the linear dependence of y_t on y_{t-1} and as this is the only term considered, this makes $\hat{\phi}_{1,1}$ the lag-1 sample PACF.

Looking inductively at the order-2 model, $\phi_{2,2}$ describes the additional dependency of y_t on y_{t-2} which y_{t-1} does not account for. By the definition of the PACF, this makes $\hat{\phi}_{2,2}$ a good estimator of the lag-2 PACF.

5.3 Asymptotic Properties and Uses

Under some regularity conditions, it can be shown that the lag-k sample PACF for an AR(p) model behaves well as T is sent to infinity. In particular, it can be shown that as $T \to \infty$ [3, p. 41],

- $\hat{\phi}_{p,p} \to \phi_p$
- $\hat{\phi}_{\ell,\ell} \to 0$ for all $\ell > p$
- For $\ell > p$, the asymptotic variance of $\hat{\phi}_{\ell,\ell}$ is $\frac{1}{T}$

This gives a good way for identifying when the PACF becomes "negligible". For large enough time series, the PACF follows a normal distribution with mean equal to its population mean and variance equal to $\frac{1}{T}$ as specified before. Therefore, as for the ACF, the PACF can be viewed as negligible if the value falls in the two-standard error interval $\left(-\frac{2}{\sqrt{T}}, \frac{2}{\sqrt{T}}\right)$[5, p. 8], giving us a valuable test statistic.

Using this and the fact that $\hat{\phi}_{\ell,\ell} \to 0$ for all $\ell > p$, we can deduce that for large enough data sets which follow an AR(p) model, the sample PACF will drop off to values close enough to 0 to be ignored for lags greater than p. Figure 8 shows a demonstration of this for a time series of length 10000 with $\phi_0 = 1$, $\phi_1 = 0.5$, $\phi_2 = -0.3$, $\phi_3 = 0.5$. The horizontal dotted blue lines on the plot are exactly the endpoints of the interval given above.

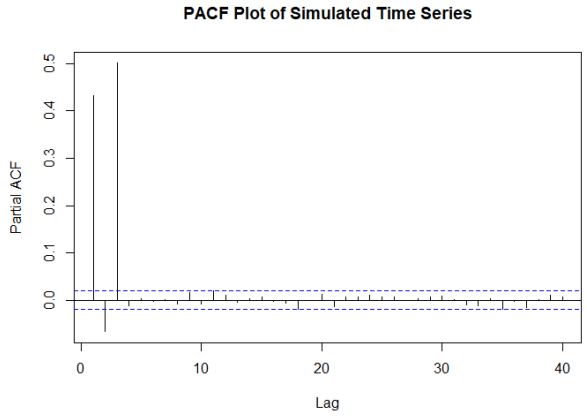

Figure 8: PACF of Synthetic AR(3) Time Series

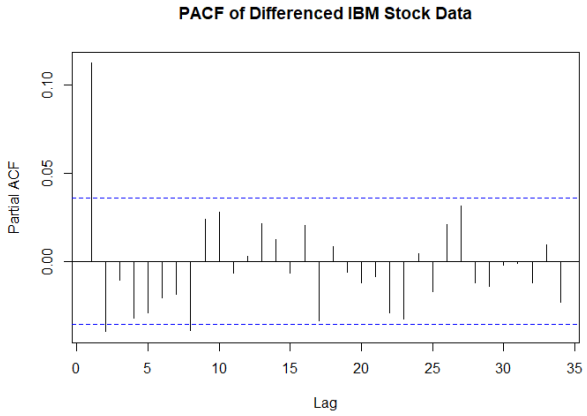

Figure 9: PACF of the Once-differenced IBM Time Series

In Figure 8, the first 3 lags show significant correlation, but all of the others beyond that fall between the blue dotted lines, and so indicate insignificant correlation. This confirms the fact that the time series generated is of order 3. In a real data set, we first look at the ACF and look for exponential decay. If we see this, we then look at the PACF in order to determine the order of the AR model.

In Figure 9, we see something very different going on. The first lag PACF is much less significant than in the simulated time series, however is still much more significant than the other lags, and sits outside of the confidence interval. The second lag is much smaller, and it sits only just outside of the interval. This means it may or may not be significant. Here, we see a good example of where the PACF gives us a decent indication of what order AR model we can use, but does not give a definitive answer. We explore tools to help us decide the order of an AR model in the Identification section.

6 Identification

When it comes to finding the model best suited to the data, there are several approaches and conditions for determining what makes a "good" model. The most simple idea when it comes to this is to look at the residuals left over after we fit a model. Assume we have a model of order p, and we have the estimates $\phi_{0,p}, \phi_{1.p}, \ldots, \phi_{p,p}, \sigma_p^2$, we can calculate the Residual Sum of Squares (RSS), and the smaller the value of this, the better a fit the model is for the data.

This method is highly ineffective, however, because the more parameters we fit, and the higher the order of the model we fit, the smaller the RSS will be. Obviously, fitting more parameters creates a better model, but this can become computationally intensive to estimate parameters for, all for decreasing gains in accuracy. This gives another potential criteria for a good model to be a model which can capture the data sufficiently well without overdoing the number of parameters in the model.

Another important tool in determining how well a model fits a set of data is Likelihood Estimation. This tells us how likely we are to observe the data under a specific model, and from this, using partial derivatives, we can obtain the Maximum Likelihood Estimates (MLE) for the parameters in our model. The better the parameters we choose, the higher the likelihood function. In the case of an autoregressive time series, finding the likelihood can be used alongside the other tools mentioned above to create a more reliable tool for judging the fit of a model.

6.1 Information Criteria

Information criteria are a very useful quantity for determining the order of a time series. When fitting a time series model to a set of data, the model will never be exact; some information will inevitably be lost. Information criteria try to quantify this loss of information, and by doing so evaluate the quality of a model.

In estimating the amount of information lost by a model, Information criteria deal with the trade-off be-

tween the goodness of fit of the model (by considering the likelihood) and the simplicity of the model (by considering the number of parameters). In other words, they compare the risk of overfitting with the risk of underfitting.

6.1.1 Akaike Information Criterion (AIC)

One such criterion is the Akaike Information criterion. It is defined as [3, p. 41]:

$$\text{AIC} = \frac{2k}{T} - \frac{2\log(\hat{L})}{T} \qquad \begin{aligned} k &= \text{Number of parameters in model} \\ \hat{L} &= \text{Maximum likelihood function value} \\ T &= \text{Total length of time series} \end{aligned}$$

This can be applied to any model we may choose to fit to the data. We look for the number of parameters which gives the smallest AIC value. For an autoregressive time series of order p, the function for AIC reduces down to

$$\text{AIC}(p) = \log(\hat{\sigma}_p^2) + \frac{2p}{T} \tag{1}$$

Using the equations for AIC, we see that it rewards a large likelihood function, and penalises us when we input too many parameters.

6.1.2 Bayesian Information Criterion (BIC)

An alternative information criterion is the Bayesian Information Criterion (BIC) [10, p. 286]:

$$\text{BIC} = k\frac{\log(T)}{T} - \frac{2\log\left(\hat{L}\right)}{T}$$

where each variable is defined in the same way as in the AIC. In comparison to the AIC, the BIC penalises more complex models, weighting the "number of parameters" term more heavily for longer time series. This makes the BIC more suitable for comparing models with larger data sets. However for smaller models, the preferred models tend to be too simple, in which case the AIC is more suitable.

6.1.3 Corrected Akaike Information Criterion (AICc)

Particularly with small sample sizes and large number of parameters we risk over-fitting our model. We introduce an adjustment accounting for the quantity of parameters [11, p. 25]:

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2k(k+1)}{T-k-1}$$

As our sample size increases the AICc will converge towards AIC.

6.2 Using the PACF

It was seen in Asymptotic Properties and Uses that the PACF of an AR(p) time series drops off to close to 0 for lag greater than p, and so using this fact and the information introduced on the AIC, we can analyse a user-generated time series, with $(\phi_0, \phi_1, \phi_2, \phi_3, \phi_4) = (1, 0.5, -0.3, 0.5, 0.2)$ and judge what the best order for this data would be.

Order	1	2	3	4	5	6	7	8
PACF	0.634	0.003	0.628	0.194	-0.002	-0.004	0.000	-0.001
AIC	4.750981	4.751172	4.246740	4.208158	4.208358	4.208552	4.208752	4.208951

Table 1: AIC and PACF Analysis of Synthetic Time Series

Looking at Table 1, it is clear that the AIC achieves a minimum at $p = 4$, and beyond order 4, the PACF does not go above a modulus of 0.004, and for a time series of length 10000, the value $\frac{2}{\sqrt{T}} = \frac{2}{100} = 0.02$, so the PACF sits comfortably within the confidence interval introduced earlier. We can also analyse the same information for the Once-differenced IBM Time Series. The analysis is shown in Figures 10 and 11 and Table 2:

Order	1	2	3	4	5	6	7	8
PACF	0.112	-0.040	-0.011	-0.032	-0.029	-0.021	-0.019	-0.039
AIC	3.921594	3.920672	3.921219	3.920842	3.920647	3.920883	3.921205	3.920326
AICc	3.922919	3.924649	3.929176	3.934109	3.940554	3.948763	3.958390	3.968150

Table 2: AIC, AICc and PACF Analysis of Once-differenced IBM Time Series

Looking at Figures 10 and 11 and Table 2, the second row of Table 2 puts the information found in Figure 9 into numbers. The third row, and Figure 10 show the AIC and how it varies over orders 1 through 8. We see significantly low values at orders 2 and 5, followed by a minimum at 8. This would make it sound like

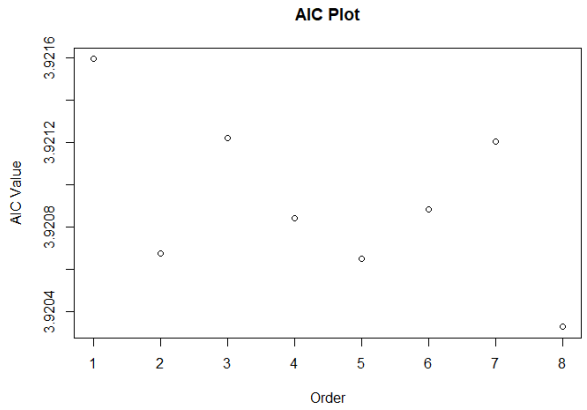

Figure 10: AIC Plot for Once-differenced IBM Time Series

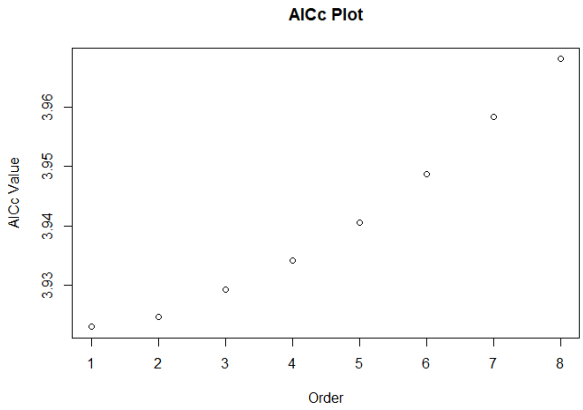

Figure 11: AICc Plot for Once-differenced IBM Time Series

an order 8 model would fit the data best. However, note that the time series is approximately 3000 in size, which makes the addition made by the corrected AIC cause a change in the shape of the plot, shown in the final row of Table 2 and in Figure 11. The AICc shows an increasing pattern, so the smaller the model, the better. This brings us to the conclusion that a model of order 2 is ideal for the data, using the low AIC and significant PACF.

7 Parameter Estimation

7.1 Least Squares Estimation

If we have a system of equations $Y = X\phi$, where Y and X are covariates, X is full rank, ϕ are parameters, we can minimise $\|Y - X\phi\|$ with the estimator $\hat{\phi} = (X^T X)^{-1} X^T Y$, which exists because if X is full rank, $X^T X$ is full rank and invertible.

Consider the case where p = 1, i.e $y_t = \phi_1 y_{t-1} + \varepsilon_t$, and suppose that $y_1, \ldots, y_n$ are known. We will assume that ε_t are i.i.d, so that the $\frac{\text{RSS}}{n - \text{rank}(X)}$ is an unbiased estimator for σ^2, where RSS is the residual sum of squares and rank(X) is the dimension of the column span of X.

$$\underbrace{\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \phi_1 \end{bmatrix}}_{\phi}$$

$$\hat{\phi} = (X^T X)^{-1} X^T Y = (\begin{bmatrix} y_1 & \dots & y_{n-1} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \end{bmatrix})^{-1} \begin{bmatrix} y_1 & \dots & y_{n-1} \end{bmatrix} \begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\hat{\phi}_1 = \frac{\sum_{i=1}^{n-1} y_i y_{i+1}}{\sum_{i=1}^{n-1} y_i^2} \qquad [9, p.\ 1]$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \left\| Y - X\hat{\phi} \right\|^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (y_{i+1} - \hat{\phi}_1 y_i)^2$$

This can be generalised for an AR(p) process but it is computationally intensive.

7.2 Maximum Likelihood Estimation

Suppose the order of our time series is 1 again, and that ε_t are independent and identically distributed. $\varepsilon_t = Y_t - \phi_1 Y_{t-1} \sim N(0, \sigma^2)$. y_1 is fixed and $y_1, \dots, y_n$ are known.

We find the likelihood by using the distribution of ε_t and then using a substitution

$$\begin{aligned} L(\phi_1, \sigma) &= P(y_1, y_2, \dots, y_n | \phi_1, \sigma) \\ &= (2\pi\sigma^2)^{\frac{-(n-1)}{2}} \exp(\frac{-1}{2\sigma^2} \sum_{t=2}^{n} \varepsilon_t^2) \\ &= (2\pi\sigma^2)^{\frac{-(n-1)}{2}} \exp(\frac{-1}{2\sigma^2} \sum_{t=2}^{n} (y_t - \phi_1 y_{t-1})^2) \end{aligned}$$

Now taking logs and, multiplying by -1:

$$-f = -\log(L(\phi_1, \sigma)) = \frac{n-1}{2}(\log(2\pi) + 2\log(\sigma)) + \frac{1}{2\sigma^2} \sum_{t=1}^{n-1} (y_{t+1} - \phi_1 y_t)^2)$$

Calculate partial derivatives and set them to 0 to find our estimators

$$\begin{aligned} -\frac{\partial f}{\partial \phi_1} &= \frac{1}{2\sigma^2} \sum_{t=1}^{n-1} (y_{t+1} - \phi_1 y_t) * -2y_t \\ -\frac{\partial f}{\partial \phi_1}|_{\hat{\sigma}, \hat{\phi}_1} = 0 &\implies \sum_{t=1}^{n-1} y_{t+1} y_t - \hat{\phi}_1 y_t^2 = 0 \\ &\implies \hat{\phi}_1 = \frac{\sum_{t=1}^{n-1} y_{t+1} y_t}{\sum_{t=1}^{n-1} y_t^2} \end{aligned}$$

$$-\frac{\partial f}{\partial \sigma} = \frac{n-1}{\sigma} - \frac{1}{\sigma^3}\sum_{t=1}^{n-1}(y_{t+1} - \phi_1 y_t)^2$$

$$-\frac{\partial f}{\partial \sigma_1}|_{\hat{\sigma},\hat{\phi}_1} = 0 \implies n-1 = \hat{\sigma}^2 \sum_{t=1}^{n-1}(y_{t+1} - \hat{\phi}_1 y_t)^2$$

$$\implies \hat{\sigma}^2 = \frac{\sum_{t=1}^{n-1}(y_{t+1} - \hat{\phi}_1 y_t)^2}{n-1}$$

Now we calculate the second partial derivatives

$$-\frac{\mathrm{d}^2 f}{\mathrm{d}\phi_1^2}|_{\hat{\phi}_1,\hat{\sigma}} = \frac{1}{2\hat{\sigma}^2}\sum_{t=1}^{n-1} y_t^2 > 0$$

$$-\frac{\mathrm{d}^2 f}{\mathrm{d}\sigma^2}|_{\hat{\phi}_1,\hat{\sigma}} = -\frac{n-1}{\hat{\sigma}^2} + \frac{3}{\hat{\sigma}^4}\sum_{t=1}^{n-1}(y_{t+1} - \hat{\phi}_1 y_t)^2$$

$$= \frac{2(n-1)}{\hat{\sigma}^2} > 0$$

As the second partial derivatives are negative, the likelihood therefore reaches a maximum at $(\phi_1, \sigma) = (\hat{\phi}_1, \hat{\sigma})$. If we were to work with autoregressive processes with a higher order, we would have to use computational methods to numerically solve our equations.

7.3 Method of Moments

In the previous two sections we only looked at estimators for AR(1) processes, because the algebra becomes very messy as the order increases, but here we can generalise to any autoregressive process.

$$\text{Suppose } y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2) \qquad t \in \mathbb{Z}$$

$(1 - \phi_1 B - \ldots - \phi_p B^p) y_t = \varepsilon_t$. We assume Y_t is stationary, which holds if all of the roots of $f(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ are outside the unit circle [12].

All autoregressive processes are invertible, [3, p. 63] which means that Y_t is a linear sum of white noise. By linearity of the expectation, this implies that $\mathrm{E}(y_t) = 0$.

We will enforce stationarity on our time series, and together with invertibility, this will help with proving some results later on.

$$\mathrm{E}(y_t \varepsilon_t) = \mathrm{E}(\varepsilon_t \sum_{k=1}^{p} \phi_k y_{t-k}) + \mathrm{E}(\varepsilon_t^2) = \mathrm{E}(\varepsilon_t)\mathrm{E}(\sum_{k=1}^{p} \phi_k y_{t-k}) + \mathrm{Var}(\varepsilon_t) - \mathrm{E}(\varepsilon_t)^2 = \sigma^2$$

$$\implies \mathrm{E}(y_t^2) = \mathrm{E}[(\sum_{k=1}^{p} \phi_k y_{t-k} + \varepsilon_t) y_t] = \mathrm{E}[\sum_{k=1}^{p} \phi_k y_{t-k} y_t + \varepsilon_t y_t]$$

$$= \sum_{k=1}^{p} \phi_k \mathrm{E}[y_{t-k} Y_t] + \mathrm{E}[\varepsilon_t y_t] = \sum_{k=1}^{p} \phi_k \gamma_k + \sigma^2$$

Now we use that $\mathrm{E}[y_t] = 0$ to conclude that $\mathrm{Var}(y_t) = \mathrm{E}[y_t^2] - \mathrm{E}[y_t]^2 = \sum_{k=1}^{p} \phi_k \gamma(k) + \sigma^2$.

Now if we define our estimators for the autocovariance $\hat{\gamma}_l$, the autocorrelation $\hat{\rho}_l$, we obtain a relationship between the following parameters estimators.

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2 \qquad \hat{\gamma}_\ell = \frac{1}{n} \sum_{i=\ell+1}^{n} (y_i - \bar{y})(y_{i-k} - \bar{y}) \qquad \hat{\rho}_\ell = \frac{\hat{\gamma}_l}{\hat{\gamma}_0}$$

$$\hat{\sigma}^2 = S^2 - \sum_{k=1}^{p} \phi_k \hat{\gamma}_k \tag{1}$$

7.3.1 Yule-Walker

We will use a similar method as before to derive the Yule-Walker equations and our MOM (Method of Moments) estimators.

$$y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + \varepsilon_t$$

$$\implies y_t y_{t-j} = \sum_{k=1}^{p} \phi_k y_{t-k} y_{t-j} + \varepsilon_t y_{t-j}$$

Then take expectations on both sides

$$\mathrm{E}[y_t y_{t-j}] = \sum_{k=1}^{p} \phi_k \mathrm{E}[y_{t-k} y_{t-j}] + \mathrm{E}[\varepsilon_t y_{t-j}]$$

Use that $\mathrm{E}[y_t y_{t+k}] = \mathrm{Cov}(y_t, y_{t+k}) + \mathrm{E}[y_{t+k}]\mathrm{E}[y_t] = \gamma_k$, $\mathrm{E}[\varepsilon_t y_{t-j}] = 0$ and divide by γ_0:

$$\gamma_j = \sum_{k=1}^{n} \phi_k \gamma_{k-j}$$

$$\implies \rho_j = \sum_{k=1}^{n} \phi_k \rho_{k-j}$$

This is a system of equations if we send j from 1 to p:

$$\underbrace{\begin{bmatrix} 1 & \rho_1 & \dots & \rho_{p-1} \\ \rho_1 & 1 & \dots & \rho_{p-2} \\ \vdots & \vdots & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \dots & 1 \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}}_{\phi} = \underbrace{\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix}}_{\rho}$$

Note that we used that $\rho_k = \rho_{-k}$, so that $R = (y_{ij})_{\rho x \rho}, y_{ij} = \rho_{|i-j|}$. The above equation $R\phi = p$ has a unique solution if we assume P is full rank, so we can obtain another relationship:

$$\hat{\phi} = \hat{R}^{-1}\hat{p} \qquad [13, p.\ 123] \tag{2}$$

Where $\hat{R}, \hat{\rho}$ contain the estimators for the autocorrelations.

We now find $\hat{\sigma}$, use that $\gamma_l = \gamma_0 \rho_l$

$$\begin{aligned} \mathrm{Cov}(y_t, y_t) &= \mathrm{Cov}(\sum_{k=1}^{p} \phi_k y_{t-k} + \varepsilon_t, y_t) \\ &= \sum_{k=1}^{p} \phi_k \mathrm{Cov}(y_{t-k}, y_t) + \mathrm{Cov}(\varepsilon_t, y_t) \\ &\implies \gamma_0 = \sum_{k=1}^{p} \phi_k \gamma_k + \sigma^2 \\ &\implies \gamma_0 = \gamma_0 \sum_{k=1}^{p} \phi_k \rho_k + \sigma^2 \\ &\implies \gamma_0 (1 - \sum_{k=1}^{p} \phi_k \rho_k) = \sigma^2 \end{aligned}$$

By replacing with parameter estimators, and using the symmetry of $\hat{R}$ and (2) we get that

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\gamma}_0 (1 - \hat{\phi}^T \hat{\rho}) \\ &= \hat{\gamma}_0 (1 - (\hat{R}^{-1}\hat{\rho})^T \hat{\rho}) \\ &= \hat{\gamma}_0 (1 - \hat{\rho}^T \hat{R}^{-1} \hat{\rho}) \qquad [13, p.\ 123] \end{aligned} \tag{3}$$

By comparing our estimators for an AR(1) process, we can see that our maximum likelihood estimators and least squares estimators are the same, but different to the method of moments estimators $\hat{\rho}_1$ and $\hat{\gamma}_0$. However all 3 estimators for ϕ_1 would be the same if $y_t = 0$. In parameter estimation, it is ideal to use Yule-Walker equations as they are easy to generalise to any auto-regressive process, as opposed to least squares estimation or maximum likelihood estimation. We can do this in R by inputting a time series and an appropriate order.

```
> yw(data,2)
$phi
[1]  0.11655129 -0.03994221

$theta
[1] 0

$sigma2
[1] 2.952522

$aicc
[1] 11842.12

$se.phi
[1] 0.01818535 0.01818535

$se.theta
[1] 0
```

Figure 12: Fit Once-differenced IBM time series to model $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$, $\texttt{\$phi} = (\hat{\phi}_1, \hat{\phi}_2), \texttt{\$sigma2} = \hat{\sigma}^2$

8 Residuals

8.1 Residual Time Series

For an AR(p) model $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$, after calculating parameter estimates $\hat{\phi}_i$ (calculated via the parameter estimation methods from Section 7, or otherwise) we construct a fitted model [3, p. 43]:

$$\hat{y}_t = \hat{\phi}_0 + \hat{\phi}_1 y_{t-1} + \hat{\phi}_2 y_{t-2} + \ldots + \hat{\phi}_p y_{t-p}$$

With the fitted model $\hat{y}_t$ it is useful to look at its residual time series $\{\hat{\varepsilon}_t\}$, defined as estimated values subtracted from the observed values:

$$\hat{\varepsilon}_t = y_t - \hat{y}_t$$

With variance of residuals [3, p. 43]:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=p+1}^{T} \hat{\varepsilon}_t^2}{T - 2p - 1}$$

We visualize in Figures 13, 14, and 15 the residuals from fitting an AR(3) model to the generic time series generated with code in Appendix A. The code used in this section can be found in Appendix B.

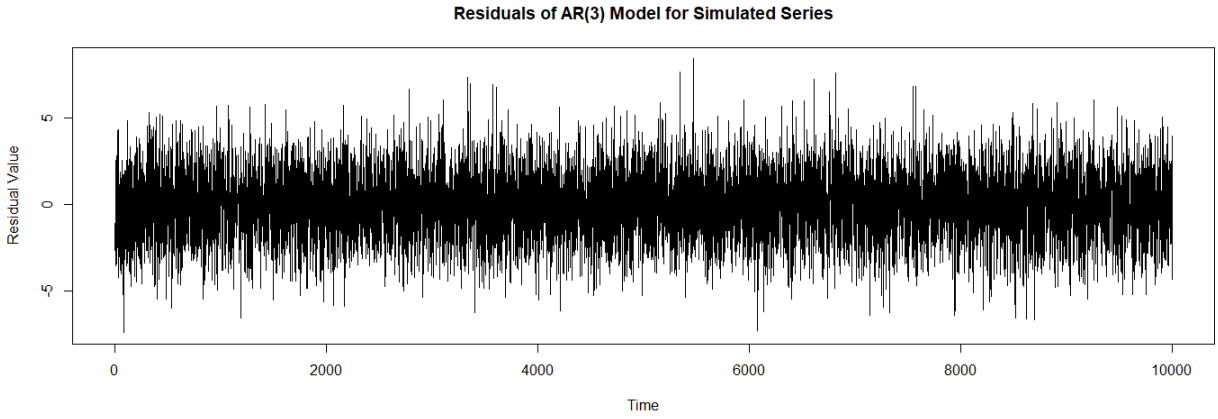

Figure 13: Residual Time Series Plot of the Residuals for Simulated Time Series

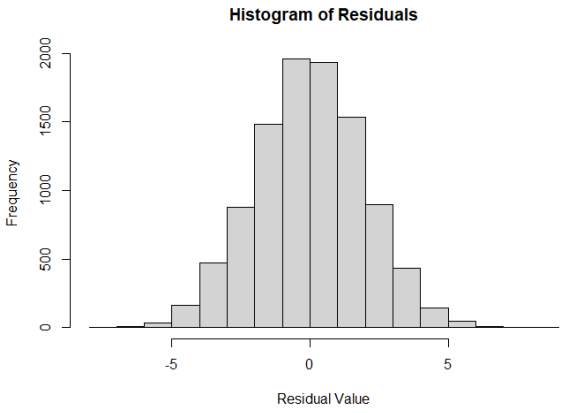

Figure 14: Histogram of Residuals for Simulated Time Series

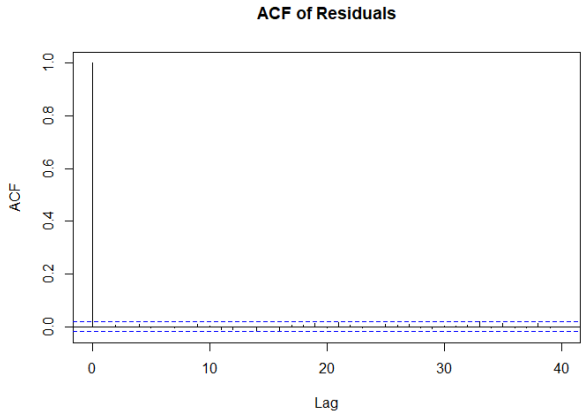

Figure 15: ACF of Residuals for Simulated Time Series

The residuals displayed in Figure 14 visually seem to follow a normal distribution well. However, to more rigorously test for deviation from a normal distribution we can run, for example, a Shapiro–Wilk or a Jarque-Bera Test to test for normality [14]. For both of these our null hypothesis H_0 is that the data (in this case, the residuals from the AR model) are normally distributed. Running the Jarque-Bera Test on our residuals with R:

```
> jarque.bera.test(ar3$residuals)

Jarque Bera Test

data:  ar3$residuals

X-squared = 0.52765, df = 2, p-value = 0.7681
```

As our p-value is significantly larger than any α significance value we could have reasonably chosen, we fail to reject the null hypothesis that the residuals are normally distributed.

8.2 Goodness of Fit

To quantify goodness of fit we use the residuals in the coefficient of determination, "R-square" for the AR(p) model [3, p. 46]:

$$R^2 = 1 - \frac{\sum_{t=p+1}^{T} \hat{\varepsilon}_t^2}{\sum_{t=p+1}^{T} (y_t - \bar{y})^2}$$

Where we define $\bar{y} = \frac{\sum_{t=p+1}^{T} y_t}{T-p}$, $\sum_{t=p+1}^{T} \hat{\varepsilon}_t^2$ is the residual sum of squares (RSS) and $\sum_{t=p+1}^{T} (y_t - \bar{y})^2$ is the total sum of squares (TSS). Our R^2 is in $[0, 1]$: it tells us the proportion of the variance of y_t explained by the chosen number of lag variables in the AR(p) model.

However, the coefficient of determination can be misleading for a non-stationary time series; the TSS grows uncontrollably as the samples size increases, and R^2 converges to 1.

As R^2 tends to increase with increases in the number of parameters, we instead introduce an adjusted R^2 that does not rely on the parameter count. Of course, this also means that this value is not in $[0, 1]$ anymore [3, p. 47]:

$$R^2_{\text{adj}} = 1 - \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_y^2}$$

σ_ε^2 and σ_y^2 representing the sample variance of the residuals and y_t respectively. When comparing between multiple AR models, we can also rely on the AIC and BIC explored earlier to compare indicators of model quality.

8.3 Testing for White Noise

To comprehensively conduct a portmanteau test [3, p. 27] for white noise we introduce a few statistics.

8.3.1 Box-Pierce and Ljung-Box Tests

The first of which is the Box-Pierce statistic [15]

$$Q_{BP}(m) = T \sum_{\ell=1}^{m} \hat{\rho}_\ell^2$$

with $\hat{\rho}_\ell$ being the sample autocorrelation with lag ℓ of the residual time series $\{\hat{\varepsilon}_t\}$, T once again being the length of the time series.

The modified version of this, is the Ljung-Box statistic [16]:

$$Q_{LB}(m) = T(T+2)\sum_{\ell=1}^{m}\frac{\hat{\rho}_\ell^2}{T-\ell}$$

Provided that the residual time series are indeed independently identically distributed, $Q_{BP}(m)$ and $Q_{LB}(m)$ converge to a χ^2 distribution with m degrees of freedom [15, 16].

Our hypotheses for a statistical test with significance α are [3, p. 27]

$$H_0 \colon \rho_1 = \cdots = \rho_m = 0$$

$$H_a \colon \rho_i \neq 0, \text{for some } i \in \{1, \ldots, m\}$$

In each test if $Q_{BP}(m)$ or $Q_{LB}(m) > \chi^2_\alpha$ we reject H_0; i.e. the evidence suggests an absence of white noise.

We can demonstrate these tests in R, again with the residuals from the simulated AR(3) model (created with the R code in Appendix A):

```
> Box.test(ar3$residuals, type="Ljung-Box")

Box-Ljung test

data:  ar3$residuals

X-squared = 0.0080019, df = 1, p-value = 0.9287

> Box.test(ar3$residuals, type="Box-Pierce")

Box-Pierce test

data:  ar3$residuals

X-squared = 0.0079995, df = 1, p-value = 0.9287
```

In both instances the p-values are significantly higher than any reasonable significance level; the data do not suggest the residuals deviate from white noise. Additionally, the ACF of residuals in Figure 15 display a lack of correlation.

8.3.2 Durbin–Watson Test

In simple AR(1) case, we can identify autocorrelation in its residual time series with the Durbin–Watson statistic [17]

$$DW = \frac{\sum_{t=2}^{T}\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2}$$

Our null and alternate hypotheses for an α significance level test are simpler: $H_0 : \rho = 0$ and $H_a : \rho \neq 0$. This statistic varies between 0 and 4; using significance tables or computational test in R we can find d_L and d_U; with a DW under or above which we reject H_0. We can also more specifically test for positive or negative correlation depending if $DW < d_L$ or $DW > d_U$ respectively.

8.3.3 Testing on IBM Data

We continue our analysis of the Once-differenced IBM time series. Using an AR(2) model on the data, we produce similar plots of residuals as with the AR(3) simulated time series (Figures 16, 17, and 18).

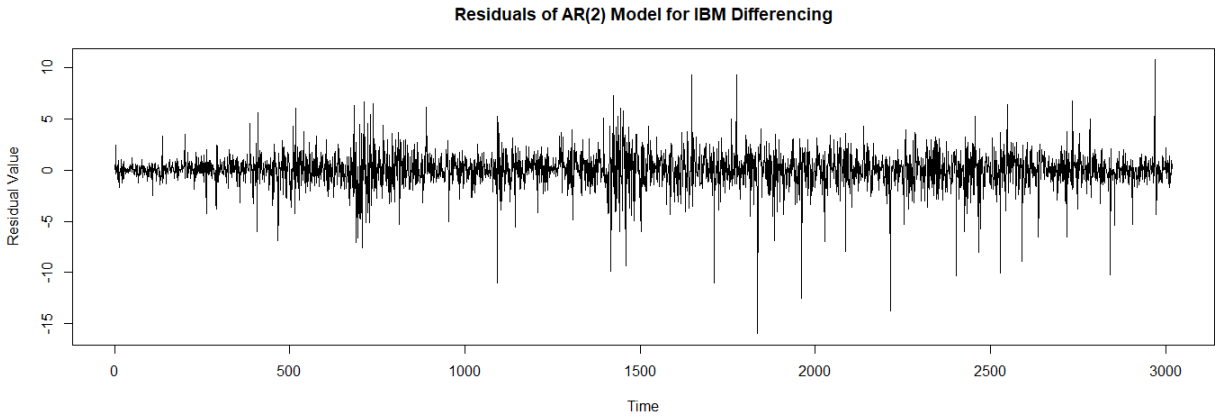

Figure 16: Residual Time Series Plot of the Residuals for Once-differenced IBM Time Series

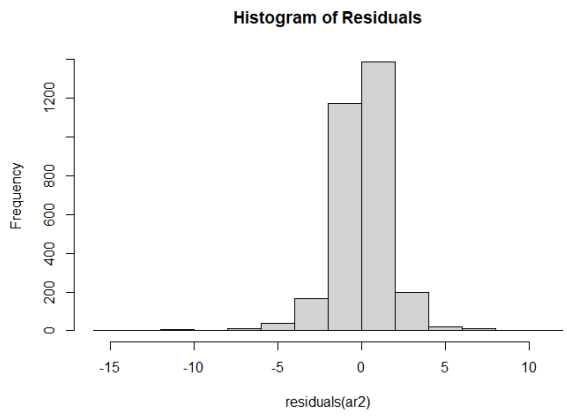

Figure 17: Histogram of Residuals for Once-differenced IBM Time Series

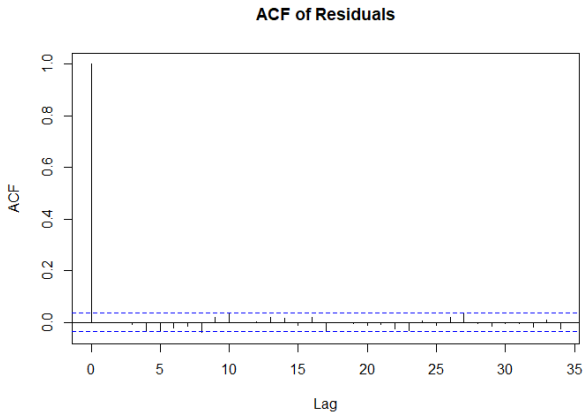

Figure 18: ACF of Residuals for Once-differenced IBM Time Series

Implementing the same statistical tests we used earlier: conducting the Jarque-Bera and Box-Pierce tests:

```
> IBM_resid <- ar2$residuals[!is.na(ar2$residuals)] # remove NA's
> jarque.bera.test(IBM_resid)
Jarque Bera Test
data:  IBM_resid
X-squared = 13588, df = 2, p-value < 2.2e-16
> Box.test(ar2$residuals, type="Ljung-Box")
Box-Ljung test
data:  ar2$residuals
X-squared = 0.00039673, df = 1, p-value = 0.9841
```

The p-value in the Jarque-Bera test is below most reasonable signifigance level; there is evidence suggesting that the residuals for the AR(2) model do not follow a normal distribution. However, the p-value generated from the Ljung-Box test is high; it does not suggest a non-random residual distribution. Therefore, there are certain limitations to using an autoregressive model for data like the Once-differenced IBM time series.

9 Model Extensions

In this section we will overview alternative models and model extensions of time series.

9.1 Moving Average (MA) Models

Considered the counterpart to the AR model, Moving Average models form another very common approach to modelling time series. An MA model predicts the value of the quantity in question as a linear combination of a white noise sequence. The general form of an MA(p) time series is [18, 41]:

$$y_t = \theta_0 + \varepsilon_t - \sum_{i=1}^{p} \theta_i \varepsilon_{t-i} \qquad \varepsilon_i \sim N(0, \sigma^2)$$

9.1.1 Properties of MA Models

Stationarity - As MA models comprise of a linear combination of a white noise sequence, all of their moments are finite and time invariant, therefore are always weakly stationary, with:

$$E(y_t) = \theta_0, \qquad \text{Var}(y_t) = (1 + \sum_{i=1}^{p} \theta_i^2)\sigma^2$$

Invertibility - One interesting property of MA models is that they can be written as infinite order AR time series, under specific conditions. By repeated substitutions, we can derive the following form for an MA(1) time series:

$$y_t = \phi_0 + \theta_1 y_{t-1} + \theta_1^2 y_{t-2} + ... + \varepsilon_t$$

This time series only converges for $|\theta_1| < 1$. In general the same can be done for any order p, resulting in a condition $|\theta_i| < 1$ for all $i = 1, ..., p$ [18, 43]. An MA time series that satisfies this condition is said to be *invertible*. This condition ensures that any given MA model has a unique ACF.

Autocorrelation - Just as with an AR model the PACF can be used to identify the order, so can the ACF be used to determine the order of an MA model. If $\rho(p) \neq 0$ and $\rho(\ell) = 0$ for all $\ell > p$, then the order of the MA time series is p. This is because the white noise variable ε_t is independent of ε_{t-p} in an MA(p) model.

9.2 Autoregressive Moving Average (ARMA/ARIMA) Models

9.2.1 Motivation

Combining ideas from the two previously discussed models, we can form a more complex model known as the Autoregressive Moving Average model. Often we are faced with systems that depend on a series of unobserved shocks as well as it's own prior behaviour, in which case we rely on ARMA models. This can help compactify the model and reduce the number of parameters involved.

9.2.2 General Form

The general Form for an ARMA model is [3, 58]:

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

9.2.3 Identification

Identifying the order of an ARMA(p,q) can be quite challenging - The ACF and PACF do not suffice. Instead the extended autocorrelation (EACF) function can be used, as proposed by Tsay and Tiao [19]. This method is significantly more complex, but the general idea is to firstly obtain a consistent estimate of the AR component by method of iterated Least Squares, then to derive the order of the MA component using the ACF.

9.2.4 Autoregressive Integrated Moving Average (ARIMA) Model

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) mode, where the given time series is non-stationary. An ARIMA model introduces a differencing step to eliminate any seasonal/trend component of the function [3, 67-68]. This method of differencing is discussed in section 9.3.

9.3 Trend, Seasonality and Differencing

In most of this report, we have considered stationary time series, where the mean is constant. One possible extension is to model the mean as a function of time instead.

9.3.1 Trend

A time series may exhibit long term change in the mean, in which case we say there exists a *trend*. An example of a trend is $y_t = \mu_t + \varepsilon_t$, where $\mu_t = \beta_0 + \beta_1 t$ and ε_t is the random error with $E(\varepsilon_t) = 0$. Then, $E(y_t) = \mu_t = \beta_0 + \beta_1 t$ and there exists a linear trend.

9.3.2 Seasonality

A time series may also possess periodic or cyclical changes in the mean, in which case it is known as a *seasonal time series*. Seasonality can be identified via a time plot or an autocorrelogram if there exists high spikes at lag multiple of the periodicity p [3, p. 74]. An example of seasonality is $y_t = \mu_t + \varepsilon_t$, where $\mu_t = E(y_t) = sin(2\pi t)$, hence y_t is a seasonal time series of periodicity 1.

9.3.3 Classical Decomposition

In general, the *classical decomposition model* takes into account a combination of trend and seasonality in data [13, p. 20]:

$$y_t = m_t + s_t + \varepsilon_t$$

m_t is the *trend component*, a function that accounts for the slow, long-term change in the mean

s_t is the *seasonal component*, a function of known periodicity that accounts for periodic fluctuations

ε_t is the *random noise component*, a random and stationary residual.

9.3.4 Differencing to Achieve Stationarity

To analyse such a model, it is convenient to temporarily transform it into a stationary model. Once all analysis is conducted on the stationary model, one can then revert back to the initial model. To achieve stationarity, the trend and seasonality components therefore need to be extracted, to be only left with the stationary residuals.

One way to do this is to use *differencing*. This method consists of transforming the original data series $\{y_t\}_{t=1}^T$ into a new time series $\{w_t\}_{t=1}^T$ to achieve stationarity. A general *s-order differencing* is defined by [18, p. 21,75]:

$$w_t = (1 - B)^s y_t$$

where B is the back-shift operator.

For instance, first order differencing is $w_t = (1 - B)y_t = y_t - y_{t-1}$, and second order differencing gives $w_t = (1 - B)^2 y_t = (1 - B)y_t - (1 - B)y_{t-1} = y_t - 2y_{t-1} + y_{t-2}$. Differencing can be applied directly to the time series y_t multiple times until it becomes stationary. It can also be used to extract a polynomial trend component. Indeed, for $y_t = m_t + \varepsilon_t$, $m_t = \sum_{j=0}^n \lambda_j t^j$ a polynomial of order n and ε_t stationary with mean zero, a n-th order differencing gives the stationary time series [13, p. 25]: $(1-B)^n y_t = n!\lambda_n + (1-B)^n \varepsilon_t$.

Similarly, for the seasonal component, one can use *seasonal differencing*. For a seasonal time series of periodicity d, this is defined as [13, p. 28]:

$$w_t = (1 - B^d)y_t = y_t - y_{t-d}$$

For instance, 12 term seasonal differencing can be used for monthly data. Under seasonal differencing, the seasonal component is eliminated. Indeed, for the classical decomposition model $y_t = m_t + s_t + \varepsilon_t$ with seasonal component of periodicty d, we have [13, p. 28]:

$$(1 - B^d)y_t = (m_t - m_{t-d}) + (s_t - s_{t-d}) + (\varepsilon_t - \varepsilon_{t-d}) = (m_t - m_{t-d}) + (\varepsilon_t - \varepsilon_{t-d})$$

where the seasonality component is eliminated due to its periodicity. The trend component can then be extracted using differencing again.

If the seasonal component is of minor importance, it can also be ignored completely in the inital model and removed, a method called *seasonal adjustment* [3, p. 72]. Note however that a more standard approach to treat a time series with a seasonal fluctuation in the mean is to fit a seasonal autoregressive integrated

moving average (SARIMA) model, an ARIMA model (see section 9.2.4) which takes into account a seasonal component [13, p. 180].

All of this is based on the assumption that the model has stationary residuals, which is not always the case. Therefore, as a preliminary step, the time series has to be transformed. For instance, if the seasonal component is directly proportional to the mean and the time series is in fact of the multiplicative form $y_t = m_t s_t \varepsilon_t$, then a log transformation yields an additive form $\ln(y_t) = \ln(m_t) + \ln(s_t) + \ln(\varepsilon_t)$ where the model now has stationary residuals [18, p. 15].

9.4 Differencing IBM data

Looking at the time plot of the original IBM time series, in Figure 19, a trend component can be observed. Therefore, the theory presented in section 9.3.4 suggests that differencing can be applied to revert to a stationary time series. From visual inspection of Figure 20, we can see that indeed the mean of the IBM time series after first order differencing appears to be constant.

This can be checked by fitting an ARIMA model (see section 9.2.4) to the IBM data series using the `auto.arima` function in R. This shows that the order of the ARIMA model is indeed (2,1,0), that is, an AR(2) model after first order differencing.

```
> auto.arima(IBM$Low)
Series: IBM$Low
ARIMA(2,1,0)

Coefficients:
         ar1      ar2
      0.1165  -0.0398
s.e.  0.0182   0.0182
```

In Figure 21, we can also look at the characteristic roots of the AR(2) model fitted to the Once-differenced IBM time series. Both roots have magnitude 0.199 (3 s.f), they are both inside the unit circle which confirms that the IBM data series is indeed stationary (see section 4.3.3) when first order differencing is applied.

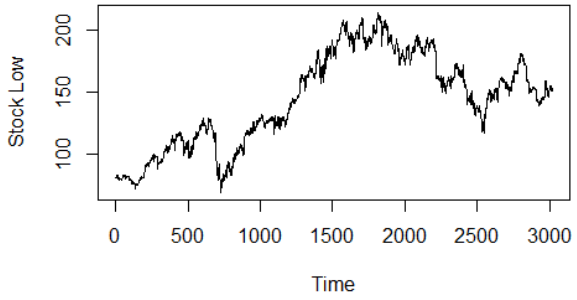

Figure 19: Time Plot of IBM Time Series

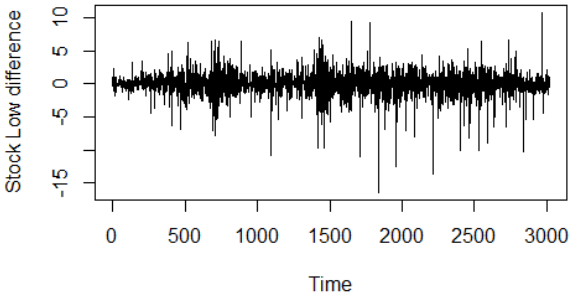

Figure 20: Time Plot of Once-differenced IBM Time Series

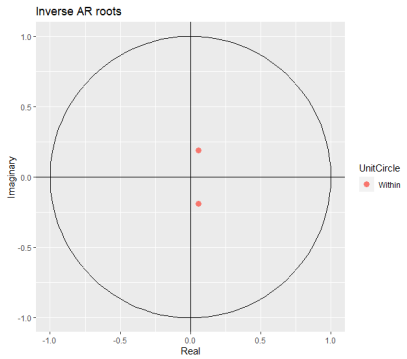

Figure 21: Characteristic Roots of Once-differenced AR model fitted on IBM Time Series

10 Forecasting

10.1 Motivation

One important goal of time series analysis is to be able to use the prior information of available observations, to predict the time series' value at some future time. This can help in sales forecasting, weather forecasting, inventory studies.

In order to achieve this, suppose that a time series $\{y_t\}_{t=1}^{T}$ is observed. Define the time index h as the *forecast origin*, such that $h < T$. That is all values $\{y_1, y_2, ..., y_h\}$ are already observed. Define also ℓ as the *forecast horizon*, where $\ell \geq 1$. Then, one can forecast the value of y_{h+l} using the observed time series values $\{y_t\}_{t=1}^{h}$. The forecast of y_{h+l} is denoted as $\hat{y_h}(l)$. There are several methods to derive $\hat{y_h}(l)$ and some are discussed in this section.

10.2 Box-Jenkins procedure

As we are mainly interested in autoregressive models in this report, a relevant method for forecasting is the Box-Jenkins procedure. This method involves fitting a mixed autoregressive integrated moving average (ARIMA) model to a given time series (see subsection 9.2.4) and then forecasting values using this model. In this section we will consider AR models instead of ARIMA as a simplification.

The Box-Jenkins forecasting procedure consists of three main steps [18, p. 90]:

1. Model identification and selection

The first step is to analyse the data in order to select the most appropriate AR model (or ARIMA model in the more general case). Several properties of the data have to be checked:

Stationarity As seen previously in the ACF section, a stationary AR(p) model should display an autocorrelogram with a mixture of exponential decay and/or damped sinusoidal oscillations. Observing no decay to zero or very slow decay therefore indicates a lack of stationarity. In this case, differencing may be used to revert to a stationary time series, as seen in section 9.3.

Trends and Seasonality If trends and seasonality are observed, then transformations, differencing, seasonal adjustement, seasonal differencing can be used to achieve stationarity as detailed in section 9.3.

Order identification The order of an AR model can be deduced from the behaviour of the ACF, the PACF or information criterion as mentioned in section 6.

2. Parameter Estimation

Once an appropriate model is selected, parameters of the AR model are estimated using methods described in section 7. This gives us the estimated equation needed for forecasts later on.

3. Diagnostic checking

This last step involves checking that the estimated model found in the previous steps satisfies the statistical properties of an AR model. One main diagnostic check is to look at residuals. Indeed, for an AR(p) model, the residuals should follow a white noise time series as mentioned in section 1.3.2, that is they must be randomly and normally distributed. To check for this, hypothesis tests and visual inspection of relevant plots can be used as explored in section 8.

If the chosen AR model is found to be inadequate in this last step, then one has to start at step 1 again to find a better AR model for the data.

Once an appropriate AR model has been found and estimated, it can be used to forecast time series values. The most direct approach to derive forecasts is to use the AR model's estimated equation and the minimum squared error (MSE) loss function. This is explored in the subsequent sections.

10.3 1-Step Ahead Forecast

Consider the AR(p) model:

$$y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t$$

where ε_t is a white noise time series with mean zero and variance σ^2.

Let F_h be the collection of information known at the forecast origin h. The objective in forecasting is to find the value $\hat{y}_h(l)$ that best predicts y_{h+l} given F_h. One way to derive the value of such optimal forecast is to minimise the Mean Square Error (MSE) between the true value y_{h+l} and its forecast $\hat{y}_h(l)$, that is [3, p. 47]:

$$MSE = E[(y_{h+l} - \hat{y}_h(l))^2 | F_h]$$

As a general result, when given a random variable Y that is predicted by a function h of random variables $X_1, X_2, ..., X_n$, one can show that the MSE $\mathrm{E}[(Y - h(X_1, ..., X_n))^2]$ is minimised for $h(X_1, ..., X_n) = \mathrm{E}[Y|X_1, ..., X_n]$ [20, p. 220].

In our forecasting scenario, treating y_{h+l} as Y, $\hat{y}_h(l)$ as the function of random variables predicting Y $h(X_1, ..., X_n)$ and F_h as $X_1, ..., X_n$, $\mathrm{E}[(y_{h+l} - \hat{y}_h(l))^2 | F_h]$ is therefore minimised for $\hat{y}_h(l) = \mathrm{E}(y_{h+l}|F_h)$.

Alternatively, a more direct derivation of this result for the 1-step ahead forecast for an AR(p) model is shown in Appendix C.

Thus, for the 1 step-ahead forecast, using linearity of expectation and that by independence of the error terms at times h and $h+1$, we obtain [3, p. 47]:

$$\hat{y}_h(1) = \mathrm{E}[y_{h+1}|F_h] = \mathrm{E}[(\phi_0 + \phi_1 y_h + ... + \phi_p y_{h+1-p} + \varepsilon_{h+1})|F_h] = \phi_0 + \sum_{i=1}^{p} \phi_i y_{h+1-i}$$

10.4 Multi-Step Ahead Forecast

10.4.1 2-Step Forecast

Now consider the equation for y_{h+2}:

$$y_{h+2} = \phi_0 + \phi_1 y_{h+1} + ... + \phi_p y_{h+2-p} + \varepsilon_{h+2}$$

In order to calculate this forecast we rely on the previous 1-step forecast (as information at time h+1 is not available). This gives [3, p. 48]:

$$\hat{y_h}(2) = \mathrm{E}[y_{h+2}|F_h] = \phi_0 + \phi_1 \hat{y_h}(1) + \phi_2 y_h + ... + \phi_p y_{h+2-p}$$

10.4.2 ℓ - Step Forecast

We repeat the process in the case of a 3-step forecast, 4-step, etc, in a recursive manner. In general, for an ℓ-step forecast [3, p. 48]:

$$\hat{y_h}(\ell) = \phi_0 + \sum_{i=1}^{\ell-1} \phi_i \hat{y_h}(\ell - i) + \sum_{i=\ell}^{p} \phi_i y_{h+\ell-i}$$

10.4.3 Forecasts of Once-differenced Time Series

Suppose that a time series $\{y_t\}_{t=1}^T$, after first order differencing, is converted to a stationary time series $\{w_t\}_{t=1}^T$, where $w_t = y_t - y_{t-1}$. Using the results derived in the sections above, one can find forecasts $\hat{w}_h(1), \hat{w}_h(2), ..., \hat{w}_h(l)$. In order to revert back to the original time series and find the ℓ-step ahead forecast $\hat{y}_h(l)$, one can simply notice that:

$$\hat{w}_h(1) = \mathrm{E}[w_{h+1}|F_h] = \mathrm{E}[(y_{h+1} - y_h)|F_h] = \hat{y}_h(1) - y_h \tag{4}$$

$$\hat{w}_h(l) = \mathrm{E}[w_{h+l}|F_h] = \mathrm{E}[(y_{h+l} - y_{h+l-1})|F_h] = \hat{y}_h(l) - \hat{y}_h(l-1) \tag{5}$$

$$\implies \sum_{k=1}^{n} \hat{w}_h(k) = \hat{y}_h(n) - \hat{y}_h(n-1) + \hat{y}_h(n-1) - ... - y_h = \hat{y}_h(n) - y_h \tag{6}$$

Using equation (6), we therefore obtain:

$$\begin{aligned}
\hat{y}_h(1) &= y_h + \hat{w}_h(1) \\
\hat{y}_h(2) &= y_h + \hat{w}_h(1) + \hat{w}_h(2) \\
&\vdots \\
\hat{y}_h(l) &= y_h + \sum_{k=1}^{l} \hat{w}_h(k)
\end{aligned}$$

10.4.4 Forecast Error

The error resulting from the forecast would be [21, p. 160]:

$$e_h(\ell) = y_{h+\ell} - \hat{y_h}(\ell) = \varepsilon_{h+\ell} + \sum_{i=1}^{\ell-1} \psi_i \varepsilon_{h+\ell-i}$$

And by the independence of the shocks, we get [21, p. 160]:

$$\mathrm{Var}(e_h(\ell)) = \left(1 + \sum_{i=1}^{\ell-1} \psi_i^2\right) \sigma_a^2$$

This result corroborates the intuitive idea that the longer the forecast, the more uncertain the prediction. For this reason forecasting is best used for short term predictions, where the model is constantly being updated as new observations are being collected.

10.4.5 Asymptotic Results

Given a stationary AR(p) model, one can show that $\hat{y_h}(\ell)$ converges to the unconditional mean $E(y_t)$ as $\ell \to \infty$. This is known as mean reversion [3, p. 66].

10.5 Prediction limits

It is important to know the precision of our forecasts. The prediction limits give a confidence interval for the value of $y_{h+\ell}$ centred around the forecast.

In an AR(p) model, if the error terms ε_t are Gaussian, then the forecast error $e_h(\ell) = y_{h+\ell} - \hat{y_h}(\ell)$ derived in section 10.4.4 also follows a normal distribution. A $(1-\alpha)100\%$ confidence interval for $y_{h+\ell}$ can therefore be constructed, where $Var(e_h(\ell))$ is estimated from the given data [20, p. 203]:

$$P\left[\left|\frac{y_{h+\ell}-\hat{y_h}(\ell)}{\sqrt{\text{Var}(e_h(\ell))}}\right| < z_{1-\alpha/2}\right] = 1-\alpha \implies \hat{y_h}(\ell) \pm z_{1-\alpha/2}\sqrt{\text{Var}(e_h(\ell))}$$

For a 95% confidence interval, this is $\hat{y_h}(l) \pm 1.96\sqrt{\text{Var}(e_h(\ell))}$, where 1.96 is often rounded up to 2.

10.6 Forecasting IBM data

We conclude this section by applying the discussed theory of forecasting on the IBM time series. We obtain the plots shown in Figures 22 and 23.

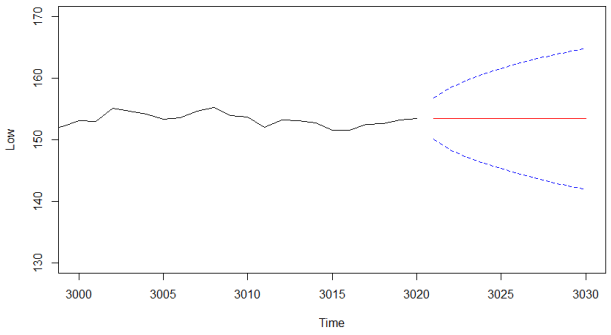

Figure 22: 10-Step Ahead Forecast of IBM Time Series

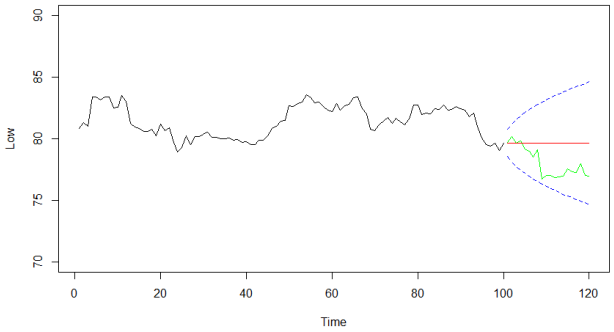

Figure 23: In-Sample Analysis: 20-Step Ahead Forecast of First 100 values of IBM Time Series

In Figure 22, the once-differenced forecast has been reverted, by summing, to get back the predicted value of the stock. The prediction limits have been identified in the plot by the dashed blue lines. The effects of mean reversion can clearly be seen where the forecast flattens out at around 153. This renders the forecasting methods useless for longer predictions.

In Figure 23, we have performed an In-Sample analysis of the first 100 values of the IBM time series, to examine the accuracy of our predictions. As we forecast values further ahead, the actual values deviate more from the predicted values; This is a consequence of the fact that longer forecasts result in errors with greater variance, as discussed in 10.4.4. They still, however, remain within the prediction limits, as expected. One idea to retain smaller prediction limits, is to continuously update the model as new observations are made, which would also avoid the effects of mean reversion.

11 Alternative Forecasting Methods

Box Jenkins is one example of a forecasting procedure, where we fit an ARIMA to estimate future observations. We could instead fit many other different models to find the best fit of past values of a time series.

11.1 Exponential Smoothing

A natural model to assume, similar to an autoregressive model, is to estimate $\hat{y_h}(1)$ by a weighted sum of all past values:

$$\hat{y_h}(1) = c_0 y_h + c_1 y_{h-1} + c_2 y_{h-2} + ...$$

As suggested by the name, in an exponential smoothing model, the coefficients c_i are given in exponential form [18, p. 99]:

$$c_i = \alpha(1-\alpha)^i$$

This follows from the intuition that more recent observations are more relevant hence should be regarded with greater importance.

The parameter α can be determined by minimising the MSE in a similar method to the Box-Jenkins l-step forecasting procedure. Values between 0.1-0.3 are typically chosen for models that rely more on values further back in time, whereas values closer to 1 are chosen for models that rely closely on recent data [18, p. 100-101].

As this form relies on an infinite number of past observations, one can alternatively compute it recursively, with $\hat{y_1}(1)$, and [18, p. 99]:

$$\hat{y_h}(1) = \alpha y_h + (1-\alpha)\hat{y}_{h-1}(1)$$

Exponential smoothing is best used to forecast first order Moving Average time series.

11.2 Holt-Winters Forecasting Procedure

One can extend exponential smoothing to estimate forecasts for models with seasonality and trends, giving the Holt-Winters Procedure. In this method of forecasting, one recursively computes the mean term, m_t, seasonal term, s_t, and trend term, r_t, to give the classical decomposition of the time series $y_t = m_t + s_t + r_t$.

The governing equations for the recursion are [18, p. 102]:

$$m_t = \alpha(x_t - s_{t-p}) + (1-\alpha)(m_{t-1} + r_{t-1})$$
$$s_t = \beta(x_t - m_t) + (1-\beta)s_{t-p}$$
$$r_t = \gamma(m_t - m_{t-1}) + (1-\gamma)r_{t-1}$$

Where $0 < \alpha, \beta, \gamma < 1$ are constants, and p is the seasonal periodicity. One can see the the similarity to the standard exponential smoothing recursive equations.

Appendices

A Simulating a Generic Time Series

Throughout the report, we have used several synthetic time series for analysis, and here, we go through the process we used for generating these series. An example of the code primarily used is shown below. Here is a step-by-step of how we decided to simulate a generic time series in R:

1. Start by assigning the values $\phi_0, \phi_1, \ldots, \phi_p$ to the variable `param`.
2. We then remove the value corresponding to ϕ_0 because this is treated differently. We also reverse the remaining list in order for it to be more easily implemented in the following steps.
3. We assign `p` to be the number of ϕ values in `coef`, `T` to be the desired length of the time series, and `sigma` to be the desired standard deviation of the errors.
4. Use the length and standard deviation to create a pre-determined vector of residuals, `res`, to be used in the AR formula, and an empty vector `Y` where we add our time series.
5. We set the first p values of `Y` to be the corresponding residuals, as these are not accounted for by the AR formula. The rest of `Y` is populated using the usual formula for an AR time series.

```
set.seed(450)
param <- c(1 , 0.5 , -0.3 , 0.5)
coef <- rev(param[-1])
p <- length(coef)
T <- 10000
```

```r
sigma <- 2
res <- rnorm(T,0,sd=sigma)
Y <- c()
Y[1:p] <- res[1:p]
for(j in (p+1):T){Y[j] <- (param[1] + sum(Y[(j-p):(j-1)]*coef) + res[j])}
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##  method            from
##  as.zoo.data.frame zoo
```

```r
auto.arima(Y)
```

```
## Series: Y
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ar3    mean
##       0.4873  -0.2917  0.5050  3.3161
## s.e.  0.0086   0.0095  0.0086  0.0660
##
## sigma^2 estimated as 3.91:  log likelihood=-21005.73
## AIC=42021.45   AICc=42021.46   BIC=42057.5
```

B Analysing Time Series Residuals

```r
# We take our AR(3) model produced from auto.arima in code in Appendix A
ar3 <- auto.arima(Y) #arima(y, order=c(3, 0, 0))


# We plot our residuals, to identify gaussian distribution/white noise
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
ts.plot(residuals(ar3), main = "Residuals of AR(3) Model for Simulated Series")
hist(residuals(ar3), main = "Histogram of Residuals")
acf(residuals(ar3), main = "ACF of Residuals")
#plot(checkresiduals(ar3), main = "test") # alternative plot method
```

```r
library(tseries)
# H0: residuals normally distributed
jarque.bera.test(ar3$residuals)
shapiro.test(ar3$residuals)

# H1: residuals randomly distributed
Box.test(ar3$residuals, type="Ljung-Box") # use Ljung-Box test
Box.test(ar3$residuals, type="Box-Pierce") # use Box-Pierce test

# We look at the IBM stock data
IBM <- read.csv("IBM.csv")
plot.ts(diff(IBM$Low)) # Low of daily stock, differencing order 1

ar2 <- auto.arima(diff(IBM$Low))
# arima(IBM£Low, order=c(2, 1, 0)) or arima(diff(IBM£Low), order=c(2, 0, 0))

# We plot our residuals, to identify gaussian distribution/white noise
IBM_resid <- ar2$residuals[!is.na(ar2$residuals)] # remove NA's
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
ts.plot(residuals(ar2), main = "Residuals of AR(2) Model for IBM Differencing")
hist(residuals(ar2), main = "Histogram of Residuals")
acf(IBM_resid, main = "ACF of Residuals")

# H0: residuals normally distributed
IBM_resid <- ar2$residuals[!is.na(ar2$residuals)] # remove NA's
jarque.bera.test(IBM_resid)
shapiro.test(IBM_resid)

# H0: residuals randomly distributed
Box.test(ar2$residuals, type="Ljung-Box") # use Ljung-Box test
Box.test(ar2$residuals, type="Box-Pierce") # use Box-Pierce test
```

C Alternative derivation of 1-step ahead MSE forecast of AR(p)

Consider the AR(p) model again:

$$y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t$$

where ε_t is a white noise time series with mean zero and variance σ^2.

Then the 1-step ahead forecast of the AR(p) model can be derived in the following alternative way:

Using the result that an AR(p) model is invertible, it can also be written as a linear combination of the previous shocks ε_i, that is [3, p. 63]:

$$y_{h+1} = \varepsilon_{h+1} + \phi_0 + \sum_{i=1}^{p} \phi_i y_{h+1-i} = \varepsilon_{h+1} + \phi_0 + \sum_{i=1}^{\infty} \psi_i \varepsilon_{h+1-i}$$

Expressing the 1-step ahead forecast in a similar way, where the ψ^*_{1+i}'s are weights to be determined under MSE:

$$\hat{y_h}(1) = \phi_0 + \sum_{i=0}^{\infty} \psi^*_{1+i} \varepsilon_{h-i}$$

Then, applying the MSE on the forecast we obtain:

$$\begin{aligned}
\mathrm{E}[y_{h+1} - \hat{y_h}(1)]^2 &= \mathrm{E}[\varepsilon_{h+1} + \sum_{i=0}^{\infty} (\psi_{1+i} - \psi^*_{1+i})\varepsilon_{h-i}]^2 \\
&= \mathrm{E}[\sum_{i=0}^{\infty} m_i \varepsilon_{h+1-i}]^2 && \text{where } m_i = (\psi_i - \psi^*_i) \text{ for } i > 0,\ m_0 = 1 \\
&= \sum_{i=0}^{\infty} m_i \mathrm{E}[\varepsilon_{h+1-i}]^2 && \text{by linearity of expectation} \\
&= (1 + \sum_{i=0}^{\infty} (\psi_{1+i} - \psi^*_{1+i})^2)\sigma^2 && \text{as } \mathrm{E}[\varepsilon_t]^2 = \sigma^2 \quad \forall t
\end{aligned}$$

This is minimised if we set to zero $\sum_{i=0}^{\infty} (\psi_{1+i} - \psi^*_{1+i})^2 \sigma^2 = 0$ and hence $\psi^*_{1+i} = \psi_{1+i}$.

Therefore, for F_h denoting the information available at forecast origin h:

$$\begin{aligned}
\mathrm{E}[y_{h+1}|F_h] &= \mathrm{E}[\varepsilon_{h+1}|F_h] + \mathrm{E}[(\phi_0 + \psi_1 \varepsilon_h + \psi_2 \varepsilon_{h-1} + ...)|F_h] \\
&= \phi_0 + \psi_1 \varepsilon_h + \psi_2 \varepsilon_{h-1} + ... \\
&= \hat{y}_h(1)
\end{aligned}$$

that is, the value of $\hat{y}_h(1)$ that minimises the MSE is the conditional expectation of y_{h+1} given F_h.

This derivation can be generalised to an arbitrary l-step ahead forecast using similar arguments [21, p. 131].

References

[1] Li Y. *DJIA 30 Stock Time Series*;. Available from: https://www.kaggle.com/szrlee/stock-time-series-20050101-to-20171231 [Accessed: June 15th 2021].

[2] Taylor JB. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*. 1993;39:195–214. Available from: https://www.sciencedirect.com/science/article/pii/016722319390009L [Accessed: June 15th 2021].

[3] Tsay RS. *Analysis of Financial Time Series*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2005.

[4] Gary Napier. *Time Series*;. Available from:https://bookdown.org/gary_a_napier/time_series_lecture_notes/ChapterThree.html [Accessed: June 16th 2021].

[5] University of Cambridge Statistical Laboratory. *Time Series*;. Available from: http://www.statslab.cam.ac.uk/~rrw1/timeseries/t.pdf [Accessed: June 15th 2021].

[6] Hamilton JD. *Time Series Analysis*. Princeton, New Jersey: Princeton University Press; 1994.

[7] Ihaka R. *Time Series Analysis*. University of Auckland Statistics Department; 2005.

[8] Date S. *Understanding Partial Auto-Correlation*;. Available from: https://towardsdatascience.com/understanding-partial-auto-correlation-fa39271146ac [Accessed: June 15th 2021].

[9] Eshel G. *The Yule Walker Equations for the AR Coefficients*; 2021. Available from: http://www-stat.wharton.upenn.edu/~steele/Courses/956/Resource/YWSourceFiles/YW-Eshel.pdf [Accessed: June 15th 2021].

[10] Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*. New York: Springer; 2004.

[11] Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*. 2011 Jan;65(1):23–35. Available from: https://doi.org/10.1007/s00265-010-1029-6 [Accessed: June 15th 2021].

[12] Aldrich EM. *Linear Time Series Models*; 2014. Available from:

https://ealdrich.github.io/Teaching/Econ211C/LectureNotes/Unit1-ARMA/causality.html [Accessed: June 15th 2021].

[13] Brockwell PJ, Davis RA. *Introduction to Time Series and Forecasting*. 3rd ed. Springer Texts in Statistics. Switzerland: Springer International Publishing; 2016. Available from: https://doi.org/10.1007/978-3-319-29854-2 [Accessed: June 15th 2021].

[14] Yazici B, Yolacan S. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*. 2007;77(2):175–183. Available from: https://doi.org/10.1080/10629360600678310 [Accessed: June 15th 2021].

[15] Box GEP, Pierce DA. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*. 1970;65(332):1509–1526. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481180 [Accessed: June 15th 2021].

[16] Ljung GM, Box GEP. On a Measure of Lack of Fit in Time Series Models. *Biometrika*. 1978;65(2):297–303. Available from: http://www.jstor.org/stable/2335207 [Accessed: June 15th 2021].

[17] Simon L, Young D. *STAT 501: Regression Methods, Lesson 14.3 - Testing and Remedial Measures for Autocorrelation*;. Available from: https://online.stat.psu.edu/stat501/lesson/14/14.3 [Accessed: June 15th 2021].

[18] Chatfield C. *The Analysis of Time Series: Theory and Practice*. Springer; 2013.

[19] Tsay RS. Time Series Model Specification in the Presence of Outliers. *Journal of the American Statistical Association*. 1986;81(393):132–141. Available from: http://www.jstor.org/stable/2287980 [Accessed: June 15th 2021].

[20] Cryer JD, Chan KS. *Time Series Analysis with Applications in R*. 2nd ed. Springer Texts in Statistics. New York: Springer; 2008.

[21] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, Inc; 2016.