STAT 40700

Assignment 1

Brian McMahon

15463152

UCD School of Maths and Stats

University College Dublin

November 28, 2024

Question 1

Question 1. Let $\{X_t\}_{t\in\mathbb{Z}}$ be a causal autoregressive process given by $X_t = \phi X_{t-2} + W_t$ where $\{W_t\}_{t\in\mathbb{Z}} \sim \text{WN}(0, \sigma^2)$. Assume that the following sample moments were obtained after observing $X_1, \ldots, X_{200}$: $\hat{\gamma}(0) = 6$ and $\hat{\rho}(2) = 0.5$ (remember the notation $\gamma(0) = \text{Var}(X_t)$ and $\rho(2) = \text{corr}(X_t, X_{t-2})$). By using the causal assumption, find the method of moments estimates for the parameters ϕ and σ^2.

TLDR Final values are $\phi = 0.5$ and $\sigma^2 = 4.5$ shown in eq. (8) and eq. (9) respectively.

We will compute the moments estimates using the Yule-Walker equations eq. (1).

$$\begin{aligned} \gamma(h) &= \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \cdots + \phi_p\gamma(h-p), \quad h = 1, 2, \ldots, p \\ \sigma^2 &= \gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2) - \cdots - \phi_p\gamma(p) \end{aligned} \tag{1}$$

As X_t is causal autoregressive process we know that X_t is stationary and that it can be represented by eq. (2).

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \quad \text{for all } |z| \leq 1 \tag{2}$$

We can then rearrange X_t into the form of eq. (2) to obtain eq. (3)

$$\begin{aligned} X_t &= \phi X_{t-2} + W_t \\ X_t &= \phi B^2 X_t + W_t \\ X_t - \phi B^2 X_t &= W_t \\ X_t(1 - \phi B^2) &= W_t \end{aligned}$$

$$\phi(z) = 1 - \phi z^2 \tag{3}$$

As all values of ϕ are 0 except when 2 we obtain eq. (4)

$$\phi_q = \begin{cases} 0 & \text{for } q \neq 2 \\ \phi & \text{for } q = 2 \end{cases} \tag{4}$$

This allows us to re-write the Yule-Walker equations as eq. (5)

$$\begin{aligned} \gamma(h) &= \phi\gamma(h-2) \\ \sigma^2 &= \gamma(0) - \phi\gamma(2) \end{aligned} \tag{5}$$

We then set $h = 2$ to give eq. (6). This will be useful later on.

$$\begin{aligned} \gamma(2) &= \phi\gamma(0) \\ \sigma^2 &= \gamma(0) - \phi\gamma(2) \end{aligned} \tag{6}$$

We then use the definition of correlation to re-write $\rho(2)$ as eq. (7).

$$\text{Corr}(X_t, X_{t-2}) = \frac{\text{Cov}(X_t, X_{t-2})}{\sqrt{(\text{Var}(X_t))(\text{Var}(X_{t-2}))}}$$
$$\text{Corr}(X_t, X_{t-2}) = \frac{\text{Cov}(X_t, X_{t-2})}{\text{Var}(X_t)} \quad [\text{As Var}(X_t) = \text{Var}(X_{t-2})]$$
$$\text{Corr}(X_t, X_{t-2}) = \frac{\gamma(2)}{\text{Var}(X_t)} \quad [\text{Cov}(X_t, X_{t-2}) = \gamma(2) \quad \text{by definition}]$$

As we know that $\rho(2) = \text{Corr}(X_t, X_{t-2})$ and $\gamma(0) = \text{Var}(X_t)$ we yield:

$$\rho(2) = \frac{\gamma(2)}{\gamma(0)} \tag{7}$$

Re-arranging the first part of eq. (6) we get:

$$\phi = \frac{\gamma(2)}{\gamma(0)}$$

Then subbing into eq. (7) we get eq. (8)

$$\boxed{\phi = \rho(2) = 0.5} \tag{8}$$

This also gives us the value of $\gamma(2) = \phi\gamma(0) = 0.5(6) = 3$

Finally we substitute these values into the second part of eq. (6) to obtain the value of σ^2 eq. (9)

$$\boxed{\sigma^2 = \gamma(0) - \phi\gamma(2) = 6 - 0.5(3) = 4.5} \tag{9}$$

Question 2

Generate 2000 realisations of length $n = 300$ each of an AR(2) process with $\phi_1 = 0.2$, $\phi_2 = 0.5$, and noise variance $\sigma^2 = 2$. Find the maximum likelihood estimates (MLEs) of the three parameters in each case and store them.

A. Plot the boxplots of the MLEs and compare them to the true values.

B. Repeat part (A) using length $n = 500$. Compare the performance of the MLEs for $n = 300$ and $n = 500$. Which case produces better estimates, and how do you explain the result?

Note on code: The following code was used to generate the MLE's and plot the box plots respectively. The simulated MLE's are held in three arrays $phi1\_estimate$, $phi2\_estimate$, $sigma\_estimate$ for ϕ_1, ϕ_2, and σ^2 respectively. Also, I encountered an interesting latex code-cell error[1] but at least it looks pretty...! Finally an interesting phenomenon was encountered where the median of the estimators $(\hat{\phi}_1, \hat{\phi}_2, \hat{\sigma}^2)$ was consistently being under-estimated compared to the value used to simulate the series $(\phi_1, \phi_2, \sigma^2)$, this was explored in the Appendix section 0.0.1.

Create Timeseries and Capture MLE's

```r
library(astsa) # import lib

# Initialize vectors for MLE estimates
phi1_estimate <- numeric(0)  # Empty vector for phi1 estimates
phi2_estimate <- numeric(0)  # Empty vector for phi2 estimates
sigma_estimate <- numeric(0)  # Empty vector for sigma^2
    estimates

# Define parameters
n_sim <- 500  # Length of each time series
phi_1 <- 0.2  # AR(1) coefficient
phi_2 <- 0.5  # AR(2) coefficient
sigma2 = 2
sigma = sqrt(sigma2)

# Create loop for 2000 iterations
for (i in 1:2000) {
  # Create the time series
  X_t <- arima.sim(n = n_sim, list(order = c(2, 0, 0), ar = c(
      phi_1, phi_2)), sd = sigma)

  # Fit the AR(2) model
  fit <- arima(X_t, order = c(2, 0, 0))

  # Extract and store the MLE estimates
  phi1_estimate <- c(phi1_estimate, fitcoef["ar1"])
  phi2_estimate <- c(phi2_estimate, fitcoef["ar2"])
  sigma_estimate <- c(sigma_estimate, fitsigma2)
}
```

[1] I've encountered a problem with the LaTeX listings package, where the $ character is not being interpreted correctly and tries to enter math mode even though it has been relieved. I believe this is a package error when using language = R as I was able to fix it by using language = Python for example. There are therefore 3 missing $ signs in the following code, in case you try to run it. There are three amendments to the code: `fitcoef["ar1"] = fit$coef["ar1"]`, `fitcoef["ar2"] = fit$coef["ar2"]`, `fitsigma2 = fit$sigma2`.

Create Box Plots for each of MLE Estimates

```r
# Plotting the boxplots
par(mfrow = c(1, 3))  # Arrange the plots in 1 row and 3 columns

# Boxplot for phi1 estimates
boxplot(phi1_estimate, main = "Boxplot of Phi1 Estimates", ylab = "Phi1", col = "lightblue")
abline(h = phi_1, col = "red", lty = 2)  # True value for Phi1 (red dashed line)

# Boxplot for phi2 estimates
boxplot(phi2_estimate, main = "Boxplot of Phi2 Estimates", ylab = "Phi2", col = "lightgreen")
abline(h = phi_2, col = "red", lty = 2)  # True value for Phi2 (red dashed line)

# Boxplot for sigma estimates
boxplot(sigma_estimate, main = "Boxplot of Sigma Estimates", ylab = "Sigma", col = "lightcoral")
abline(h = sigma2, col = "red", lty = 2)  # True value for Sigma (red dashed line)

# Reset plotting layout
par(mfrow = c(1, 1))  # Reset to default single plot layout
```

Part A

The box plots of the MLE's when $n = 300$ can be seen in fig. 1. The distribution of the MLE's: $(\hat{\phi}_1, \hat{\phi}_2, \hat{\sigma}^2)$ can be seen in respect to the simulated values $(\phi_1, \phi_2, \sigma^2)$ in red. Notice that the values $(\phi_1, \phi_2, \sigma^2)$ are all located within the inner two quartiles close to the median values of their respective MLE's. We also notice a systematic difference of the median from the true values, this bias seems to be shifting the median below the simulated value for all three variables.

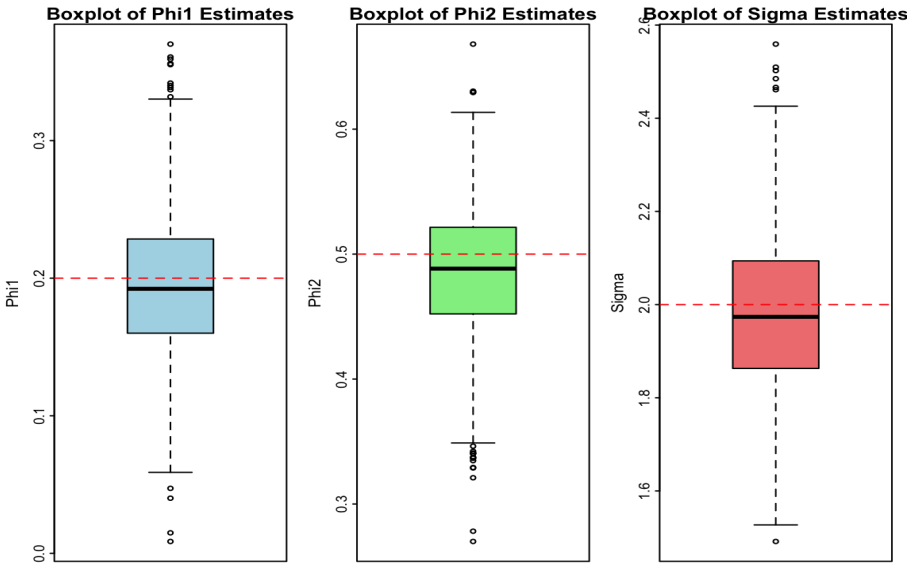

Figure 1: Box plot of estimated MLE's from 2000 simulated AR(2) processes, with time series size n=300, $\phi_1 = 0.2$, $\phi_2 = 0.5$, and noise variance $\sigma^2 = 2$. Value for simulated process shown in red.

Part B

The box plots of the MLE's when $n = 500$ can be seen in fig. 2. From visual inspection we notice that the MLE's have performed better for $n = 500$ than $n = 300$. There are two main considerations for making this statement. Firstly, the overall width of the quartiles and tails has shrunk and is closer to the value used to simulate the series. Secondly, the bias has shrunk, this can be seen when examining the median of the MLE's and seeing that they are now closer to the values used to simulate the series.

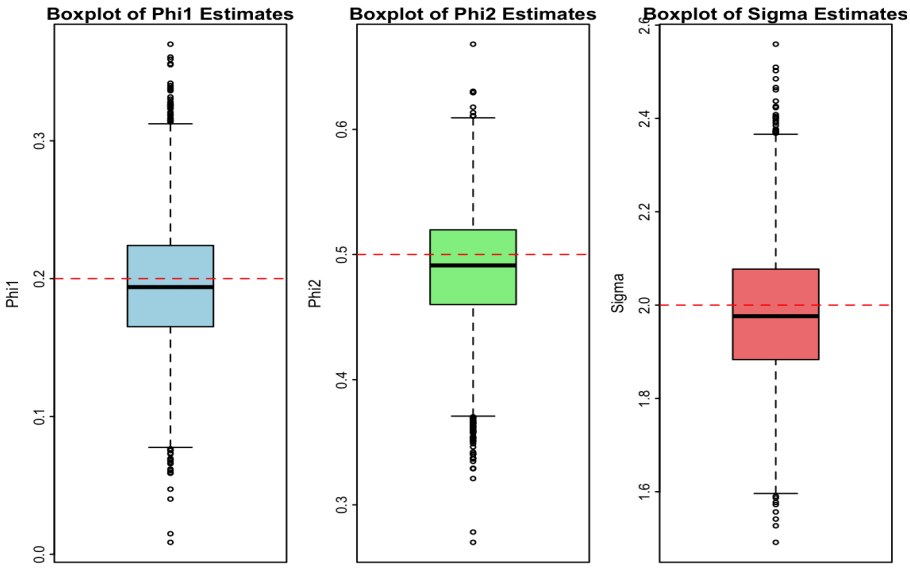

Figure 2: Box plot of estimated MLE's from 2000 simulated AR(2) processes, with time series size n=500, $\phi_1 = 0.2$, $\phi_2 = 0.5$, and noise variance $\sigma^2 = 2$. Value for simulated process shown in red.

This behaviour is shown to continue as the value of n grows. Table 1 shows that as n increases the variance decreases and the median of $(\hat{\phi}_1, \hat{\phi}_2, \hat{\sigma}^2)$ approaches the values of $(\phi_1, \phi_2, \sigma^2)$ respectively. The box plots for $n = 10000$ can be seen in Appendix A fig. 13. Finally in Appendix B we perform an investigation of increasing the number of simulated AR(2) processes section 0.0.1 and show that as the number of simulated processes increases the median approaches the bias value for the given sample size n.

Sample Size (n)	Parameter	Variance	Median
300	ϕ_1	2.679 e-03	1.94 e-01
	ϕ_2	2.513 e-03	4.894 e-01
	σ^2	2.771 e-02	1.963 e-00
500	ϕ_1	1.578 e-03	1.951 e-01
	ϕ_2	1.517 e-03	4.925 e-01
	σ^2	1.620 e-02	1.990 e-00
10000	ϕ_1	7.460 e-05	2.002 e-01
	ϕ_2	7.267 e-05	4.995 e-01
	σ^2	8.234 e-04	1.999 e-00

Table 1: Variance and Median of Parameter Estimates for Different Sample Sizes

Question 3

The time series eu15 represents the yearly sum of the GDP of the countries Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, and the UK from 1960 to 2009.

A. Does this time series appear stationary? Identify the number of differences (if any) necessary to achieve stationarity.

B. Propose an ARMA model and estimate the parameters using maximum likelihood estimation. Report the model fitting and justify your answers.

C. Comment about the statistical significance (use a significance level of 5%) of lags included in your model in (B).

D. Examine the residuals of your proposed model. Use diagnostic tools, report the relevant plots and comment on them. Is the model well fitted to the data?

E. Provide forecasts for the sum of the GDP for the years 2010, 2011, 2012, 2013, and 2014 and their respective standard errors.

Part A

TLDR 1 Differences is needed to make eu15 stationary. Figure 5 shows stationary series.

To access the stationarity of the data we use several metrics under different differencing regimes. We use ACF, PACF, Ljung-Box, and Augmented Dickey-Fuller test (ADF) to access stationarity in the first three differences, and compare results. Plots for all differencing regimes are shown Appendix A section 0.0.2, here we show the plots for only the base and chosen differencing regime. It is concluded that **one differencing** is sufficient to make the data stationary.

By visual inspection of fig. 3 we see that the time series eu15 is not stationary as the mean is clearly not constant over time. The autocorrelation function can be seen in fig. 4 where the slow decay of the ACF indicates non-stationarity. One final thing to note is the presence of a significant outlier at the end of the time series, which will impact our results later. We investigate the effects of this outlier on our used test statistics (Augmented Dickey-Fuller test and Box-Ljung test).

As differencing is clearly needed we begin by taking the first differencing and plotting the time series, ACF, and PACF in fig. 16, plots moved to appendix [2]. The differenced time series appears to have a mostly stationary mean, other that the final point, which is an outlier. The ACF quickly decays to within the statistical boundary and the PACF remains within the statistical boundaries at all lags, indicating that there is low correlation at different lags. From these plots the time series appears to become stationary after one differencing, however we also observe the Augmented Dickey-Fuller and Box-Ljung Test.

Results for augmented Dickey-Fuller (ADF) test and Ljung statistic, for each of the differences up to three, are shown in table 2, the tests are then repeated after removing the outlier. We see that the Box-Ljung test indicates that one differencing is probably sufficient in order to make our data stationary. The ADF shows that more than three differences are needed in order to have a p-value < 0.05 indicating that the data is stationary. However there is evidence to believe that the poor performance of one differencing is due to the outlier at the final position in our time series. Chan highlights the significance of outliers and their various types [1], while Frances and Haldrup [2] identify the weakness of the Dickey-Fuller test in the presence of additive outliers. As

[2] Note that some plots have been moved to appendix as strategy for replacing points has been developed to deal with the outlier.

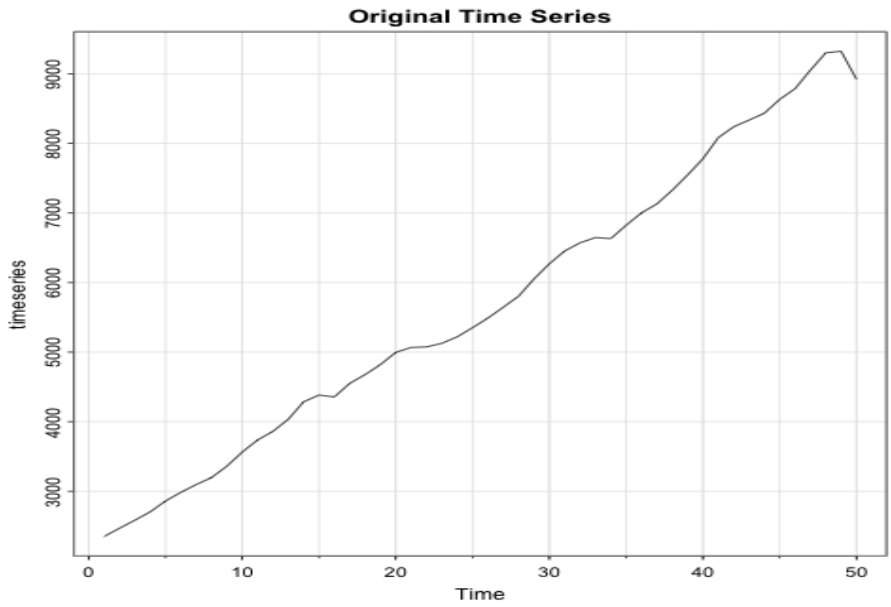

Figure 3: Plot of original time series eu15 with no differencing. Increasing mean indicates non-stationarity.

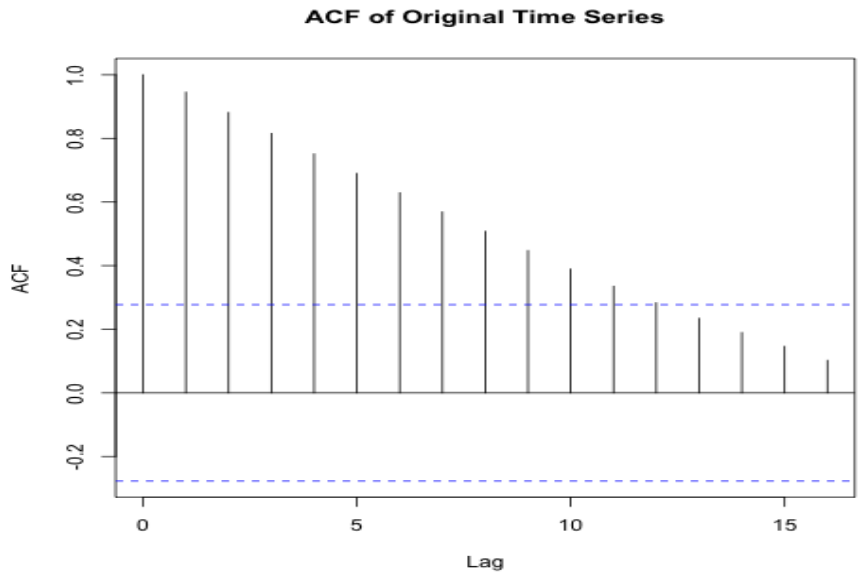

Figure 4: Plot of ACF for the original time series eu15 with no differencing. Slow decay of ACF indicates non-stationarity.

excessive differencing can destroy meaningful information in the data along with introducing other trends [3] it is important for us to minimise the amount of differencing used. We therefore repeat our previous tests while removing the outlier. We see that with one differencing is sufficient after removing the outlier.

Chan highlights that a time series contaminated by a outliers Z_t or non-repetitive interventions can be viewed as an outlier-free time series X_t plus an exogenous intervention η, $Z_t = X_t + \eta$. These outliers can be caused by interruptive global events such as strikes, pandemics or economical crises. Starting in 2007 the collapse of the US sub-prime mortgage market caused a global financial crisis [4], which could be the cause of our dataset outlier. As it is in the final position of our dataset it is difficult to tell if this is a temporary change or a level shift of the time series, but for the purposes of this assignment we will devise a policy to minimise it's effects. Our goal then is to

highlight the outlier and provide a policy for its removal or replacement. Several approaches exist in time series for replacing outliers[5]. These include: **Mean**, **Median**, **Largest Order**, **Statistic, and ARMA Forecasting** replacement. For the purposes of keeping the process simple, and as we're already forecasting in section E, we will replace the outlier with the median of the stationary series[3]. i.e. when using the differenced time series we will replace the outlier with the median of the other points. Note that the impact of the economical crises also seems to be present in the year 2008 but its effects are lessened and the point remains within the confidence interval for a stationary time series, so we choose to leave the penultimate point.[4] The adjusted time series and its ACF can now be seen in fig. 5, where we have good visual evidence of stationarity after one differencing. The ACF quickly decays also indicating stationarity. Observing table 2 one can see that with the replacement strategy and one differencing the data appears to be stationary from the ADF and BL tests. Thus the number of differences needed to achieve stationarity is 1.

		P-Value		
Test	**Data**	**Original**	**Outlier Removed**	**Augmented**
ADF Test	0 Diff	0.4507	0.7659	N/A
	1 Diff	0.3592	0.01865	0.01516
	2 Diff	0.2206	0.01	0.01
	3 Diff	0.083	0.01	0.01
BL Test	0 Diff	2.2e-16	2.2e-16	N/A
	1 Diff	0.7351	0.3679	0.3556
	2 Diff	0.7658	0.01843	0.006874
	3 Diff	0.003744	0.0001131	7.096e-05

Table 2: Comparison of p-values for Augmented Dickey-Fuller (ADF) and Box-Ljung (BL) tests for original TS, outlier removed, and outlier adjusted

Determine Number of Differences/ Test Stationarity

```
#Load Data
eu15 <- c(2349.11, ... , 8922.44)

#Define a Function for Repetitive Plots
test_stationarity <- function(timeseries) {
  tsplot(timeseries)
  acf(timeseries)
  pacf(timeseries)
  print(adf.test(timeseries))
  print(Box.test(timeseries, lag = 20, type = "Ljung-Box"))
}

#Difference Data
eu15_diff1 <- diff(eu15)

#Replace Outlier with Median
eu15_diff1[(length(eu15_diff1))] = median(eu15_diff1[1:(length(
    eu15_diff1) - 1)])

#Test Stationarity of Various TimeSeries
test_stationarity1(eu15_diff1)
```

[3] I was going to ask you about weather replacement was appropriate, but didn't get a chance in week 11, the demonstrators in the lab thought it seemed good if I justified it, so I went ahead with it.

[4] Even though we are attempting to remove the influence of the economical crises from our dataset we choose to leave the initial downturn point (penultimate point) as its within our confidence interval and we haven't shown that it's not just regular variance in the data.

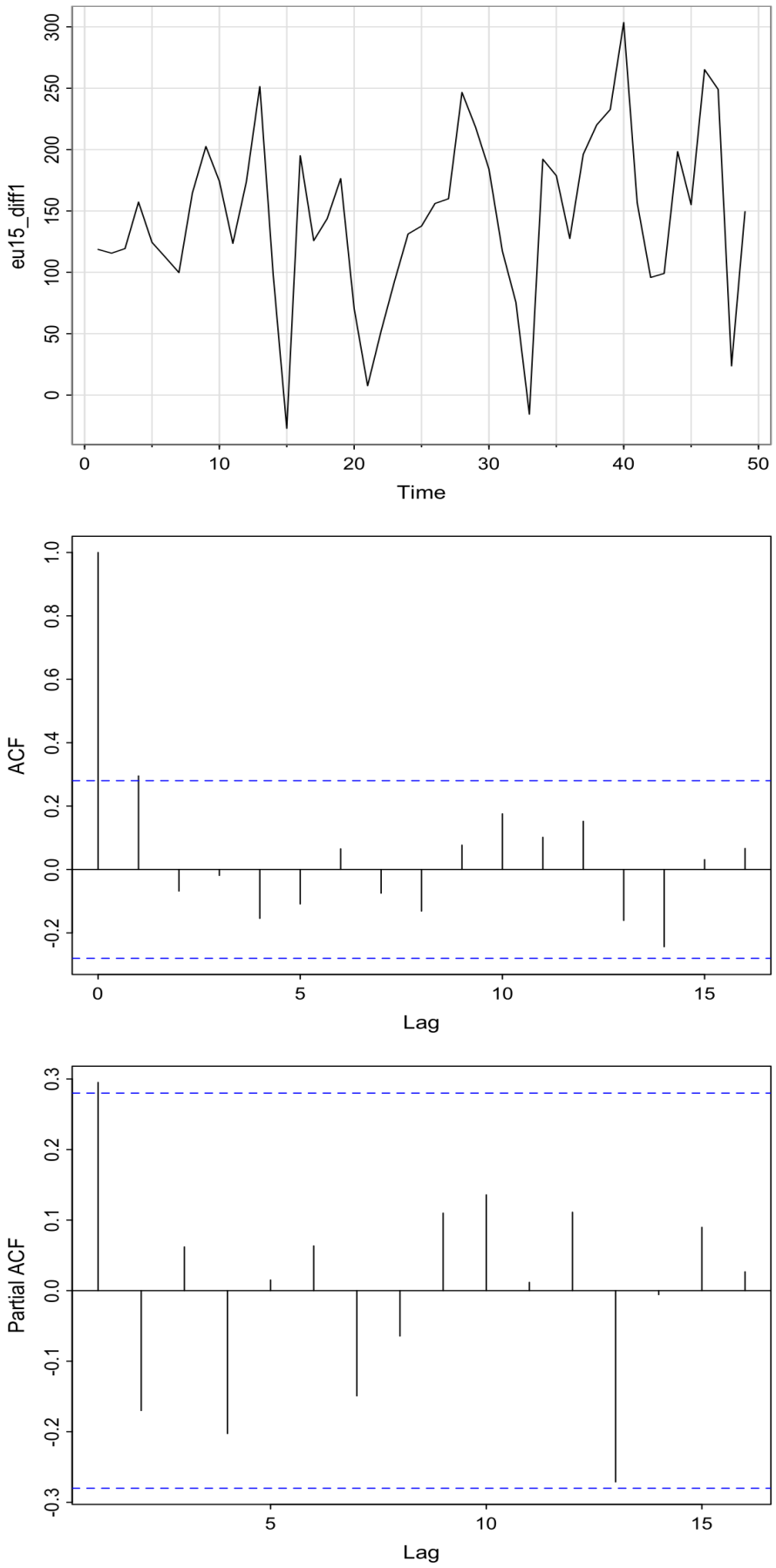

Figure 5: Plot of time series (eu15), ACF, and PACF with one differencing, with replacement of outlier in favour of median.

Part B

In this section we investigate the fitting of several different ARMA models, based on ACF, PACF, and minimising the AIC and BIC statistic. We use the augmented times series with one differencing, highlighted in the previous section as our best and most applicable stationary time series. We also investigate the original time series and outlier removed time series under the same fitting methodology, the results from this are show in appendix-A section 0.0.3.

ACF and PACF can give clues about what kind of model's will be appropriate. The ACF where lag $= 1$ is statistically significant. This indicates that this might be a moving average process MA(1)[3]. Similarly the significant PACF lags can indicate the order of the autoregressive process, from the PACF plot the first lag is significant so potentially the process is AR(1). Table 3 highlights the results of our investigation when fitting the augmented time series, with various ARMA models. Based on the principle of minimising AIC and BIC, the moving average model of order 1, MA(1), does the best at representing our data, so this is what we chose. This is also in agreement with the investigations in appendix-A for the time series with removed outlier, however the original time series is better represented with a moving average model of order 3, MA(3). The coefficients related to the MLE fitting for MA(1) are shown in table 4.

Model	df	AIC	BIC
ma1	3	554.3007	559.9761
ma2	4	555.6337	563.2010
ma3	5	557.6328	567.0919
ar1	3	556.2029	561.8784
ar2	4	556.7134	564.2807
ar3	5	558.3666	567.8257
arma11	4	555.6831	563.2504
arma21	5	557.6195	567.0786
arma31	6	559.5183	570.8692
arma12	5	557.6195	567.0786
arma13	6	559.5183	570.8692

Table 3: AIC and BIC Values for Various ARIMA Models with the associated degrees of freedom

Parameter	Estimate	Standard Error
MA(1)	0.4281	0.1552
Intercept	145.8152	13.1775

Table 4: Estimated Coefficients for the MA(1) Model

Create Several Models and Return AIC and BIC

```
#Create Various ARMA Models and Print AIC and BIC Values
ma1<-arima(eu15_diff1,order=c(0,0,1),include.mean=T)
ma2<-arima(eu15_diff1,order=c(0,0,2),include.mean=T)
ma3<-arima(eu15_diff1,order=c(0,0,3),include.mean=T)
ar1<-arima(eu15_diff1,order=c(1,0,0),include.mean=T)
ar2<-arima(eu15_diff1,order=c(2,0,0),include.mean=T)
ar3<-arima(eu15_diff1,order=c(3,0,0),include.mean=T)
arma11<-arima(eu15_diff1,order=c(1,0,1),include.mean=T)
arma21<-arima(eu15_diff1,order=c(2,0,1),include.mean=T)
arma31<-arima(eu15_diff1,order=c(3,0,1),include.mean=T)
arma12<-arima(eu15_diff1,order=c(2,0,1),include.mean=T)
arma13<-arima(eu15_diff1,order=c(3,0,1),include.mean=T)
AIC(ma1,ma2,ma3,ar1,ar2,ar3,arma11,arma21,arma31,arma12,arma13)
BIC(ma1,ma2,ma3,ar1,ar2,ar3,arma11,arma21,arma31,arma12,arma13)
```

Part C

Using our choice of the MA(1) model we only consider the first lag as the ACF for MA(q) models is zero for lags greater than q [6]. By using a T-test with the significance level of 5% we can determine if the parameter estimates reported in table 4 are significant to their error. For a two-tailed T-test the critical value is approximately 1.96 for a large number of observations relative to parameters. To investigate the significance of the 0^{th} lag we use the estimate and standard error of the intercept (μ). To investigate the significance of the 1^{st} lag we use the estimate and standard error of the first moving average parameter (MA(1)). The results of this are shown in table 5, and both the intercept and the MA(1) coefficient are shown to be significant, where $t = |\frac{Estimate}{SE}|$.

Parameter	Estimate	Standard Error (SE)	t-Statistic
MA(1) Coefficient	0.4281	0.1552	2.76
Intercept	145.8152	13.1775	11.06

Table 5: Statistical Significance of Parameters in MA(1) Model

Part D

To examine the residuals of our proposed model we make use of the SARIMA function to plot the residuals, ACF, QQ-plot, and Ljung Box statistic. We then proceed to show that the residuals of the model are as expected and the model is well fitted to the data [7]. Finally we perform the Shapiro-Wilk normality test to determine if the residuals are normally distributed.

Residuals: The residuals should look and behave like white noise $WN(0, \sigma^2)$ for a well fitted model, exhibiting mean of 0, constant variance, and uncorrelated lags. From visual inspection of the residuals plot we can see that the residuals do in fact exhibit the features associated with white noise.

ACF: For a well fit model all lags should be within the confidence interval outlined by the ACF function. This is true for our model MA(1).

QQ-Plot: For residuals the QQ-plot should contain all points within the confidence interval outlined by the gray area on the plot. All points are inside this, highlights that even for the high variance observations they are inside our confidence interval.

Ljung-Box Test: For the Ljung-Box test all p-values should be above the 0.05 confidence value for different lags. This is true for our model residuals.

Shapiro-Wilk Normality Test: We performed the Shapiro-Wilk normality test to determine if the residuals are normally distributed. The results from this test can be seen in table 6. From the p-value of 0.5633 we fail to reject the null hypothesis, that the residuals are normally distributed.

Conclusion on Residuals: We have examined the residuals plot, ACF, QQ plot, Ljung-Box plot, and the Shapiro-Wilk normality test. For all of these our model appears to meet the residual criteria for a well fit model. Thus we say that our model is fit well to our data in regards to the residuals.

Test	Value
Shapiro-Wilk W Statistic	0.97995
p-value	0.5633

Table 6: Shapiro-Wilk Normality Test for Residuals

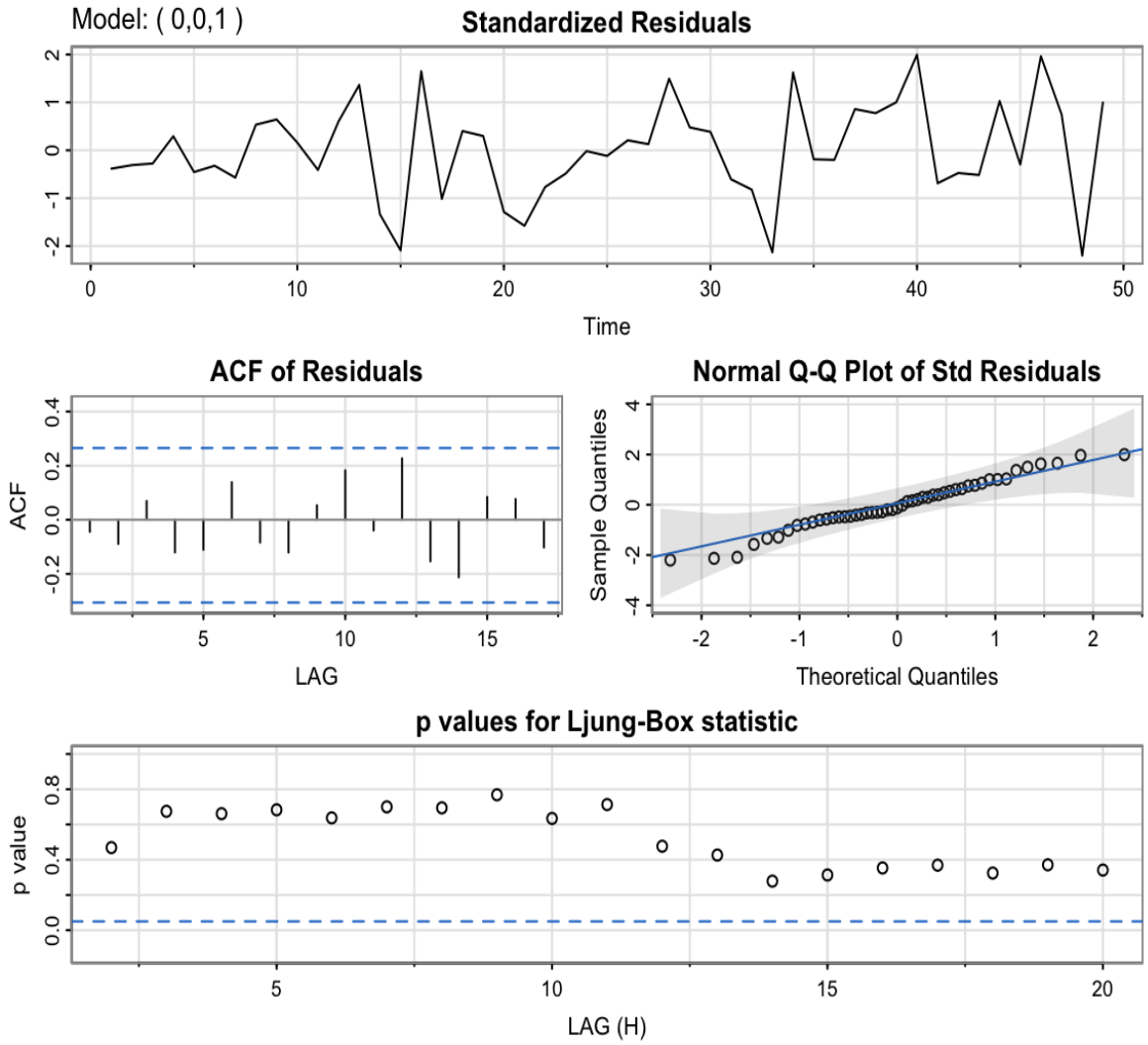

Figure 6

Get Model Params and Residuals (ma1fit substitutes ma1$fit)

```
#Create Various ARMA Models
sma1<-sarima(eu15_diff1, p=0, d=0, q=1)
print(sma1)
res1<-resid(ma1fit)
shapiro.test(res1)
```

Part E

To forecast using our model we use SARIMA package (.for). The forecasts along with their standard error (SE) are shown in table 7. We see that as the number of steps of our prediction increases so does the standard error, indicating that we are less sure as time goes on. Figure 7 shows the forecast plot along with the error quartiles in grey.

Year	2010	2011	2012	2013	2014
Forecast	8784.114	8912.677	9041.240	9169.803	9298.366
SE	92.57138	173.15261	226.70287	269.82689	306.95100

Table 7: Forecast Annual Values for 2010-2014

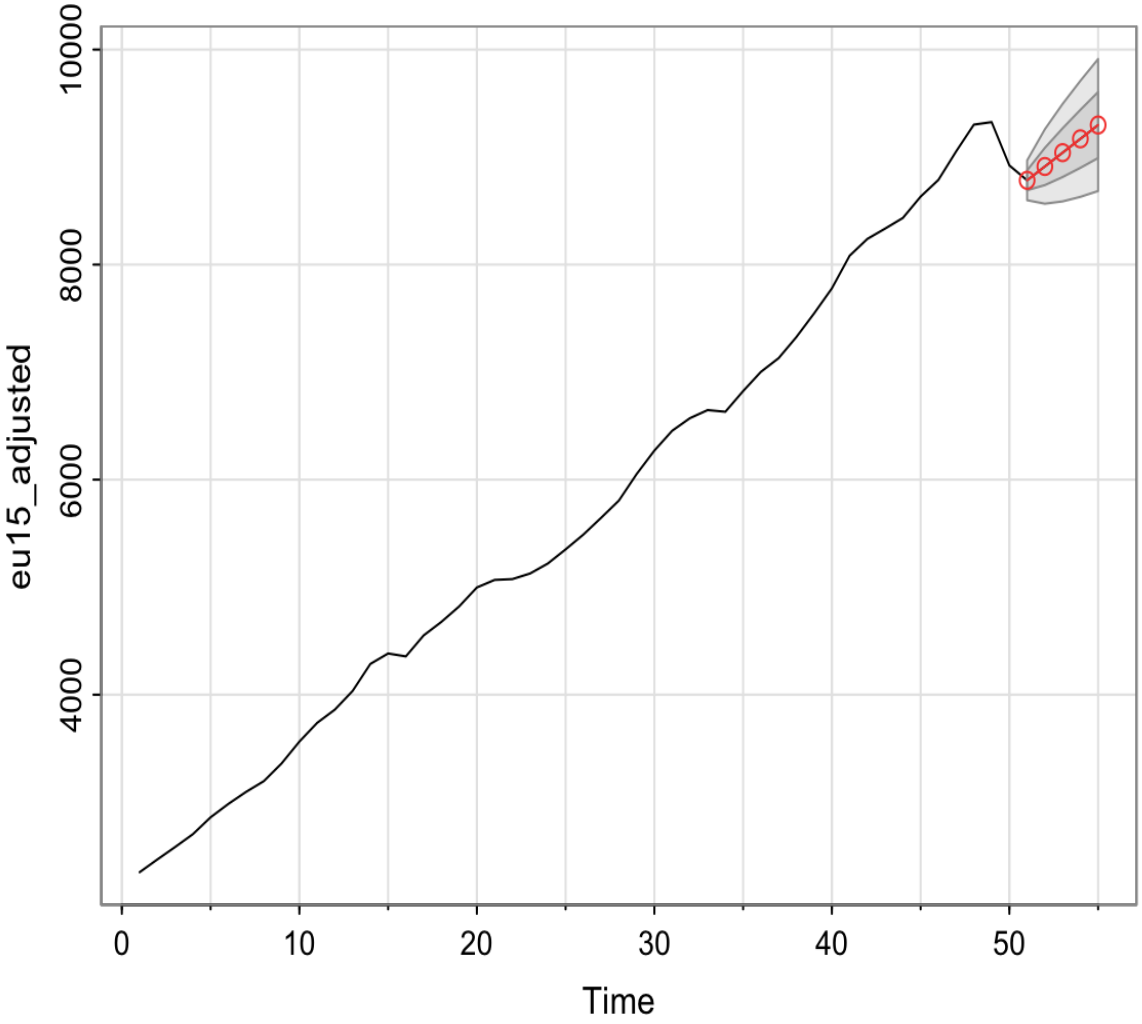

Figure 7: Plot of forecast of time series eu15 for 5 steps. Error quartiles shown in grey.

Create Forecasts with the Model

```
#Various Forecasts
eu15_adjusted<- window(eu15)
sarima.for(eu15_adjusted, n.ahead=5, p=0, d=1, q=1, plot.all=
   TRUE)
```

Question 4

Consider the time series log-income and log-consumption about the quarterly real income and real consumption expenditure respectively (both in logarithm scale and seasonally adjusted) from the UK starting in 1966:4 until 1991:2. The aim here is to investigate a possible long-term correlation between consumption and income.

(a) Are these time series $\mathrm{I}(1)$ (integrated of order 1)? Plot the time series, ACFs and perform the Augmented Dickey-Fuller test (use the significance level at 5%).

(b) Test for cointegration between log-income and log-consumption by combining Ordinary Least Squares estimation and the Augmented Dickey-Fuller test. Consider $X_t =$ log-income$_t$ and $Y_t =$ log-consumption$_t$ according to the notation considered in class. What is the conclusion of the test?

Part A

TLDR log-income and log-consumption are $\mathrm{I}(1)$ as shown in fig. 10 and fig. 11

For a non-stationary time series $\{X_t\}$ such as $\nabla^{d-1}X_t$, $\{X_t\}$ is integrated of order d if $\nabla^d X_t$ is stationary [8]. Thus for our time series' log-income and log-consumption, they are both $\mathrm{I}(1)$ if ∇log-income and ∇log-consumption are stationary respectively. Seasonal effects have already been removed as well as log transforming the data. This implies that the variance should be constant over the time series'.

It is clear from fig. 8 and fig. 9 that the original time series log-income and log-consumption are non-stationary. The time series' have a changing mean over time and ACF which decays slowly indicating non-stationarity. Figure 10 and fig. 11 indicate that the differenced time series' might be stationary. As through visual inspection the series' appear to have constant mean around 0 and similar variance throughout. The autocorrelation plots agree with this as the ACF quickly decays to within the statistical allowance. We notice that at some lags the ACF exceeds the confidence bounds at $(\pm\frac{1.96}{\sqrt{N}})$ [9]. For example at lag 14 in fig. 10 (bottom) the ACF value slightly exceeds the bound at 95%. However these high values are isolated and non-persistent. Along with the fast decaying ACF this still indicates stationarity despite the isolated breach of the confidence interval. These breaches could be caused by chance or short term variance which has not been removed by our log scaling and seasonal adjustment. As the data is real world data we accept these variances/outliers which slightly exceed the confidence bounds.

Alongside the visual methods of determining stationarity we also use the augmented Dickey-Fuller test (ADF). The ADF test for the existence of a unit root to the characteristic equation by assuming the null hypothesis of $\alpha = 1$ for $X_t = \alpha X_t - 1 + u_t$ where u_t is any stationary process (using notation from book). We know that a process with a unit root ($\alpha = 1$) is non-stationary or a less than unit root ($\alpha > 1$) non-stationary and explosive [10]. So the alternative process of $\alpha < 1$ must be true for the process to be stationary. An alternative test called the Phillips-Perron test (PP) also test for the presence of a unit root, which we use also just as another point of information.

Table 8 shows the results of these tests along with the result of the Box-Ljung statistic. The ADF and PP tests both indicate rejection of the null hypothesis, that a unit root exists, for one differencing of the time series'. Instead we accept the alternative hypothesis that no unit root exists, indicating that the data is stationary. As expected from visual inspection the original undifferenced time series' fails to reject the null hypothesis and is expected to contain a unit root, thus being non-stationary. See code for plots and tests at end of section.

Considering the time series', ACF plots, and ADF/PP tests, there is strong evidence to believe that both log-income and log-consumption are $\mathrm{I}(1)$ or integrated of order 1.

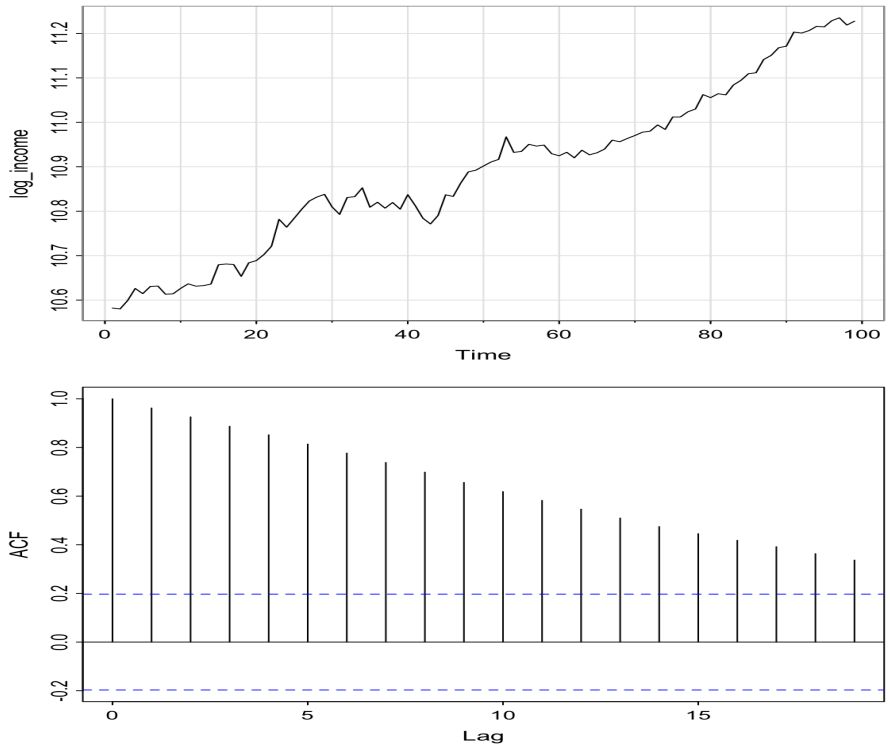

Figure 8: Time series plot (top) and ACF plot (bottom) for log-transformed income data before differencing. The series appears non-stationary, as suggested by a lack of mean reversion and significant autocorrelation.

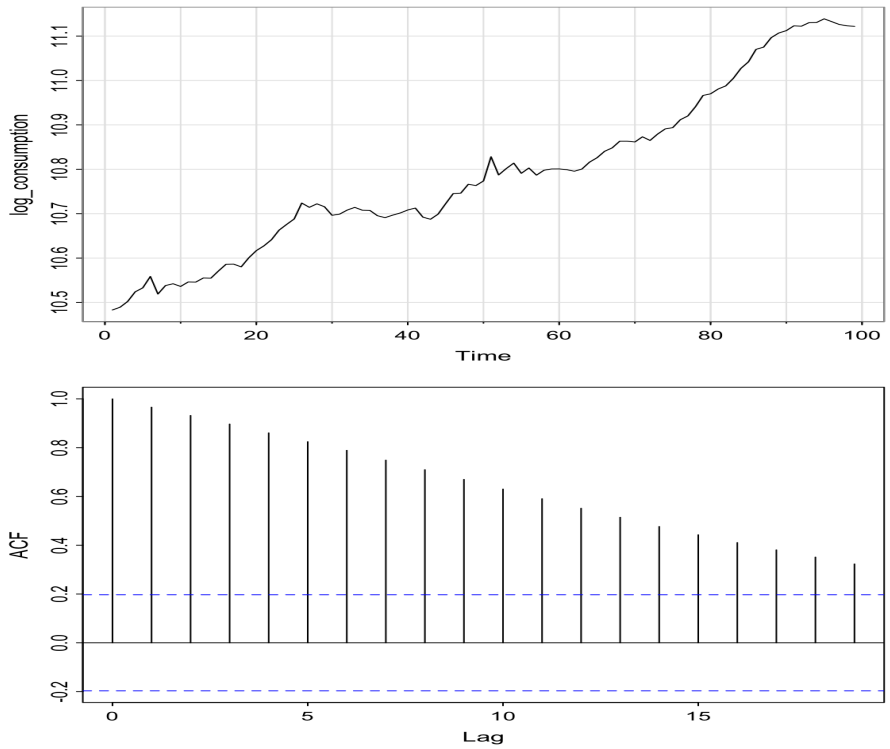

Figure 9: Time series plot (top) and ACF plot (bottom) for log-transformed consumption data before differencing. The series appears non-stationary, with a noticeable trend and significant autocorrelation.

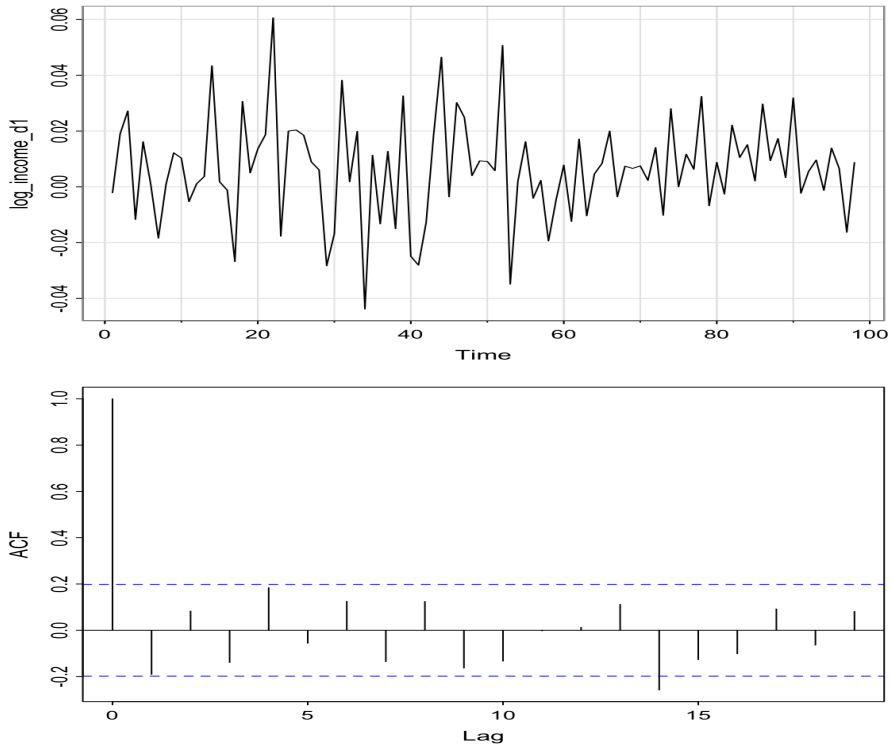

Figure 10: Time series plot (top) and ACF plot (bottom) for first-differenced log-transformed income data. The differencing removes the trend, and the series now appears stationary, with autocorrelations within the confidence bounds.

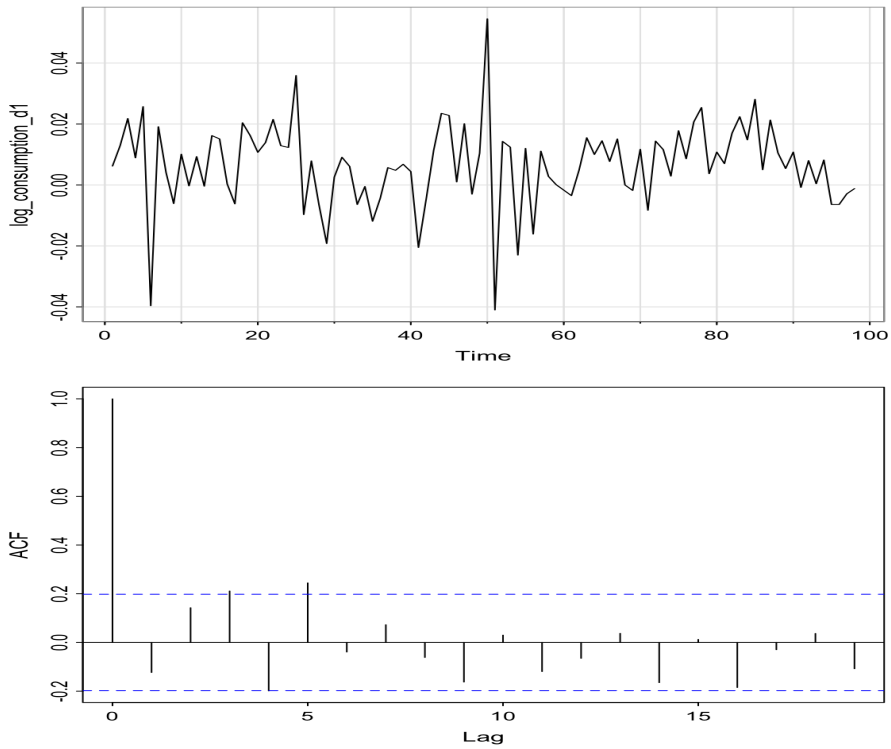

Figure 11: Time series plot (top) and ACF plot (bottom) for first-differenced log-transformed consumption data. The differencing removes the trend, and the series now appears stationary, with most autocorrelations falling within the confidence bounds.

Test	Data	P-Value	
		Original	1 Difference
Augmented Dickey-Fuller Test	log-income	0.4481	0.01334
	log-consumption	0.6387	0.04717
Phillips-Perron Test	log-income	0.482	<0.01
	log-consumption	0.7728	<0.01
Box-Ljung Test	log-income	< 2.2e-16	0.01171
	log-consumption	< 2.2e-16	0.004751

Table 8: Comparison of p-values for ADF, Box-Ljung, and Phillips-Perron Tests (Original vs. 1 Difference Data)

Create Timeseries and ACF plots, and ADF/PP Tests

```r
library(tseries) #Import lib

#Initialise arrays and first difference
lg_consumption<-c(10.4831, ... ,11.1220)
lg_income<-c(10.5821, ... ,11.2276)
lg_consumption_d1 = diff(lg_consumption)
lg_income_d1 = diff(lg_income)

#Plot TimeSeries
tsplot(lg_consumption)
tsplot(lg_income)
tsplot(lg_consumption_d1)
tsplot(lg_income_d1)

#Plot ACF
acf(lg_consumption)
acf(lg_income)
acf(lg_consumption_d1)
acf(lg_income_d1)

#Box-Ljung Test
Box.test(lg_consumption, lag = 20, type = "Ljung-Box")
Box.test(lg_income, lag = 20, type = "Ljung-Box")
Box.test(lg_consumption_d1, lag = 20, type = "Ljung-Box")
Box.test(lg_income_d1, lag = 20, type = "Ljung-Box")

#Augmented Dickey-Fuller Test
adf.test(lg_consumption)
adf.test(lg_income)
adf.test(lg_consumption_d1)
adf.test(lg_income_d1)

#Philips-Perron Test
pp.test(lg_consumption)
pp.test(lg_income)
pp.test(lg_consumption_d1)
pp.test(lg_income_d1)
```

Part B

TLDR log-income and log-consumption are not co-integrated according to ADF test and fig. 12

For two non stationary time series X_t and Y_t to be co-integrated there must exist some linear combination $Z_t = aX_t + bY_t$ where Z_t is stationary. We have shown in part A both time series log-income and log-consumptionare integrated of order one I(1). Checking that the series are non-stationary is the first step to checking their co-integration.

From the notes[8] we re-arrange eq. (10) Where $\theta = -\frac{a}{b}$

$$\begin{aligned} Z_T = aX_T + bY_t = b(Y_t) + \frac{a}{b}X_t = b(Y_t - \theta X_t) \\ Z_t = Y_t - \alpha - \theta X_t \\ Y_t = Z_t + \alpha + \theta X_t \end{aligned} \tag{10}$$

We then estimate the parameters α and θ using ordinary least squares (OLS). This allows us to compute the residuals Z_t using the estimated parameters $\hat{\alpha}$ and $\hat{\theta}$.

$$Z_T = Y_t - \hat{\alpha} - \hat{\theta}X_t$$

Finally we can determine if our residuals are stationary using the normal visual tests along with the Augmented Dickey-Fuller test. The residuals can be obtained and tested using the below R code.

Test for Stationary Residuals

```
log_consumption<-c(10.4831,...,11.1220)
log_income<-c(10.5821,...,11.2276)

ols<-lm(log_income~log_consumption)
residuals<-resid(ols)

tsplot(residuals)
acf(residuals)
adf.test(residuals)
```

By observing fig. 12 we see that the variance of the residuals appears to vary over time, and the ACF has a slow decay as the lags increase. Observing the augmented Dickey-Fuller test results table 9 we see that the test fails to reject the null hypothesis that the residuals have a unit root, i.e. are not stationary.

By observing the plots and ADF test results we have sufficient evidence to say that the time series log-income and log-consumption are **not co-integrated**.

Statistic	P-Value
Dickey-Fuller Test	0.6713

Table 9: Results of the Augmented Dickey-Fuller Test on Residuals

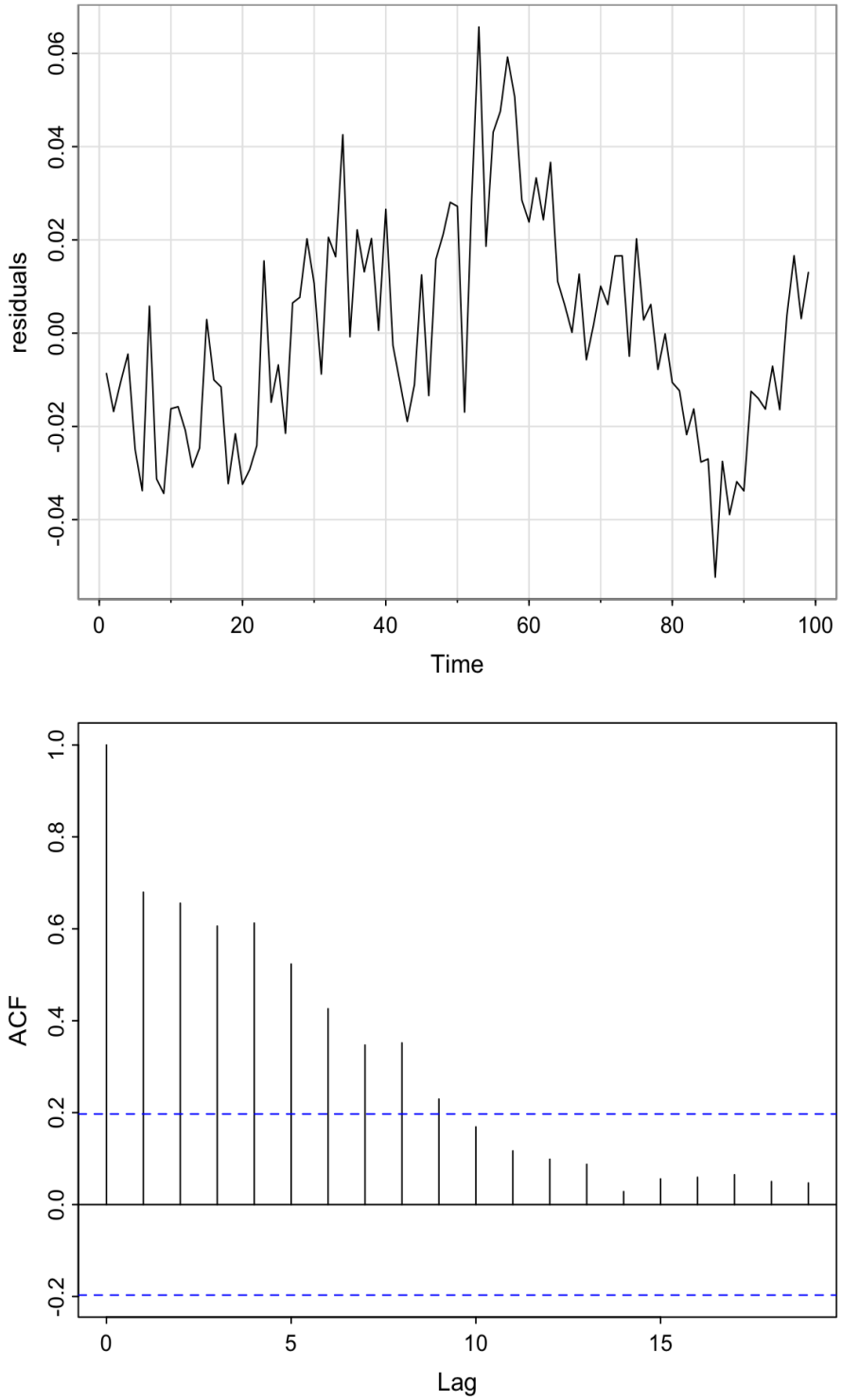

Figure 12: Plot of residuals and ACF function. Appears that co-integration is not present as the residual time series is not stationary.

References

[1] W.-S. Chan, "Understanding the effect of time series outliers on sample autocorrelations," *Test*, vol. 4, pp. 179–186, 1995.

[2] P. H. Franses and N. Haldrup, "The effects of additive outliers on tests for unit roots and cointegration," *Journal of Business & Economic Statistics*, vol. 12, no. 4, pp. 471–478, 1994.

[3] D. W. Barreto-Souza, *Slide 6 - arima and forecasting - module details 2024-25 - time series analysis stat30010/stat40700*, Accessed 22/11/24. [Online]. Available: %5Curl%7Bhttps://brightspace.ucd.ie/d2l/le/content/275957/viewContent/3134790/View%7D.

[4] D. Hodson and L. Quaglia, "European perspectives on the global financial crisis: Introduction," *JCMS: Journal of Common Market Studies*, vol. 47, no. 5, pp. 939–953, 2009.

[5] L. Appaia and S. Palraj, "On replacement of outliers and missing values in time series," *EQA-International Journal of Environmental Quality*, vol. 53, pp. 1–10, 2023.

[6] D. W. Barreto-Souza, *Slide 12 - time series - full notes - acf and pacf of arma models - time series analysis stat30010/stat40700*, Accessed 25/11/24. [Online]. Available: %5Curl%7Bhttps://brightspace.ucd.ie/d2l/le/content/275957/viewContent/3296862/View%7D.

[7] D. W. Barreto-Souza, *Slide 9 - time series - full notes - diagnostic tools - time series analysis stat30010/stat40700*, Accessed 25/11/24. [Online]. Available: %5Curl%7Bhttps://brightspace.ucd.ie/d2l/le/content/275957/viewContent/3296862/View%7D.

[8] D. W. Barreto-Souza, *Slide 14 - time series - multivariate time series and cointegration - time series analysis stat30010/stat40700*, Accessed 22/11/24. [Online]. Available: %5Curl%7Bhttps://brightspace.ucd.ie/d2l/le/content/275957/viewContent/3376143/View%7D.

[9] D. W. Barreto-Souza, *Slide 18 - time series - full notes - stationarity - time series analysis stat30010/stat40700*, Accessed 22/11/24. [Online]. Available: %5Curl%7Bhttps://brightspace.ucd.ie/d2l/le/content/275957/viewContent/3296862/View%7D.

[10] P. S. P. Cowpertwait and A. V. Metcalfe, *Introductory Time Series with R*. New York, NY: Springer, 2009, pp. 80–82, 214–216, ISBN: 9780387886978. DOI: 10.1007/978-0-387-88698-5. [Online]. Available: https://doi.org/10.1007/978-0-387-88698-5.

Appendices

Appendix A

Appendix A contains some musings around the questions.

0.0.1 Question 2 - Investigation into the bias in median estimators of our three variables.

Investigation into the bias in median estimators of our three variables. When computing the MLE's for question 2 an interesting phenomenon was noticed where the median of the estimates seemed to be consistently lower than the value used to generate the AR(2) process. We would expect for many simulations to obtain a normal distribution around the value used to simulate the dataset, so for example if $\phi_1 = 0.2$ then $\hat{\phi}_1 \sim N(0.2, \sigma_\omega^2)$. However it is clear from fig. 14 that there's an inherit bias in the estimation which is asymptotically approached by increasing the number of simulations, and reduced by increasing the size of the simulated dataset (n). We see from fig. 13 that as the size of the simulated dataset becomes large ($n = 10000$) the effect of the bias is diminished to zero.

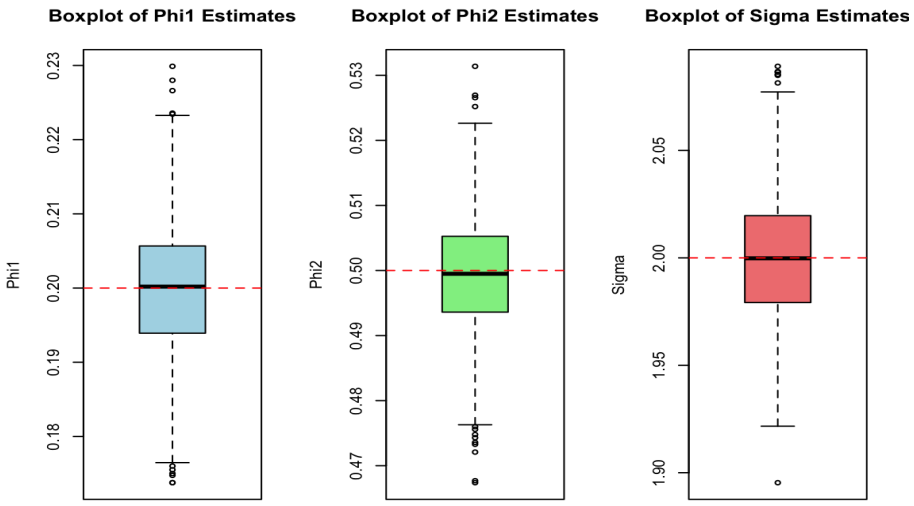

Figure 13: Box plot of estimated MLE's from 2000 simulated AR(2) processes, with time series size n=10000, $\phi_1 = 0.2$, $\phi_2 = 0.5$, and noise variance $\sigma^2 = 2$. Value used for simulated process shown in red.

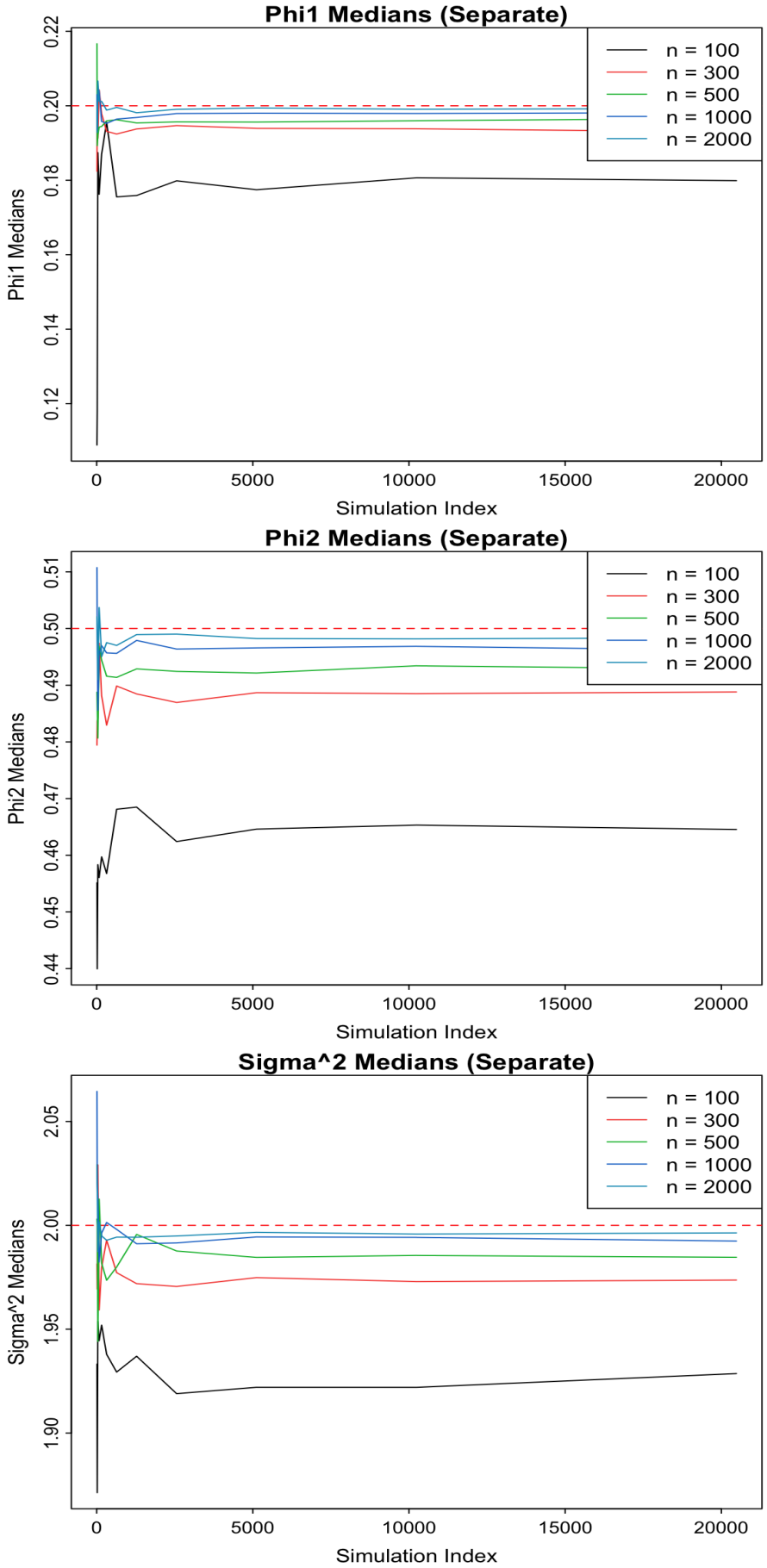

Figure 14: Plot of estimators approaching bias value. This took a long time to run XD

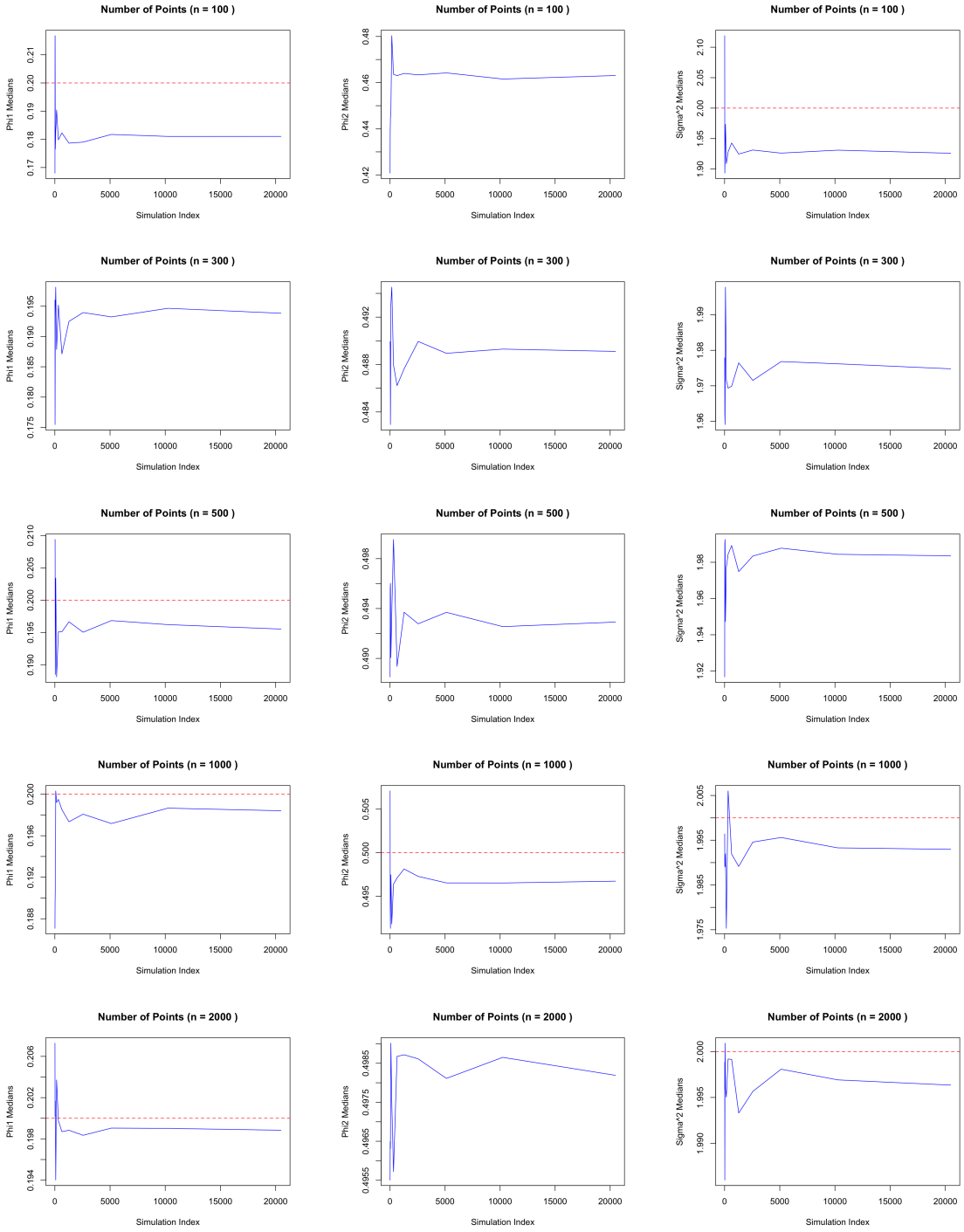

Figure 15: Grid of Images as estimators approach bias value

0.0.2 Question 3 - Investigating Stationarity of Time Series

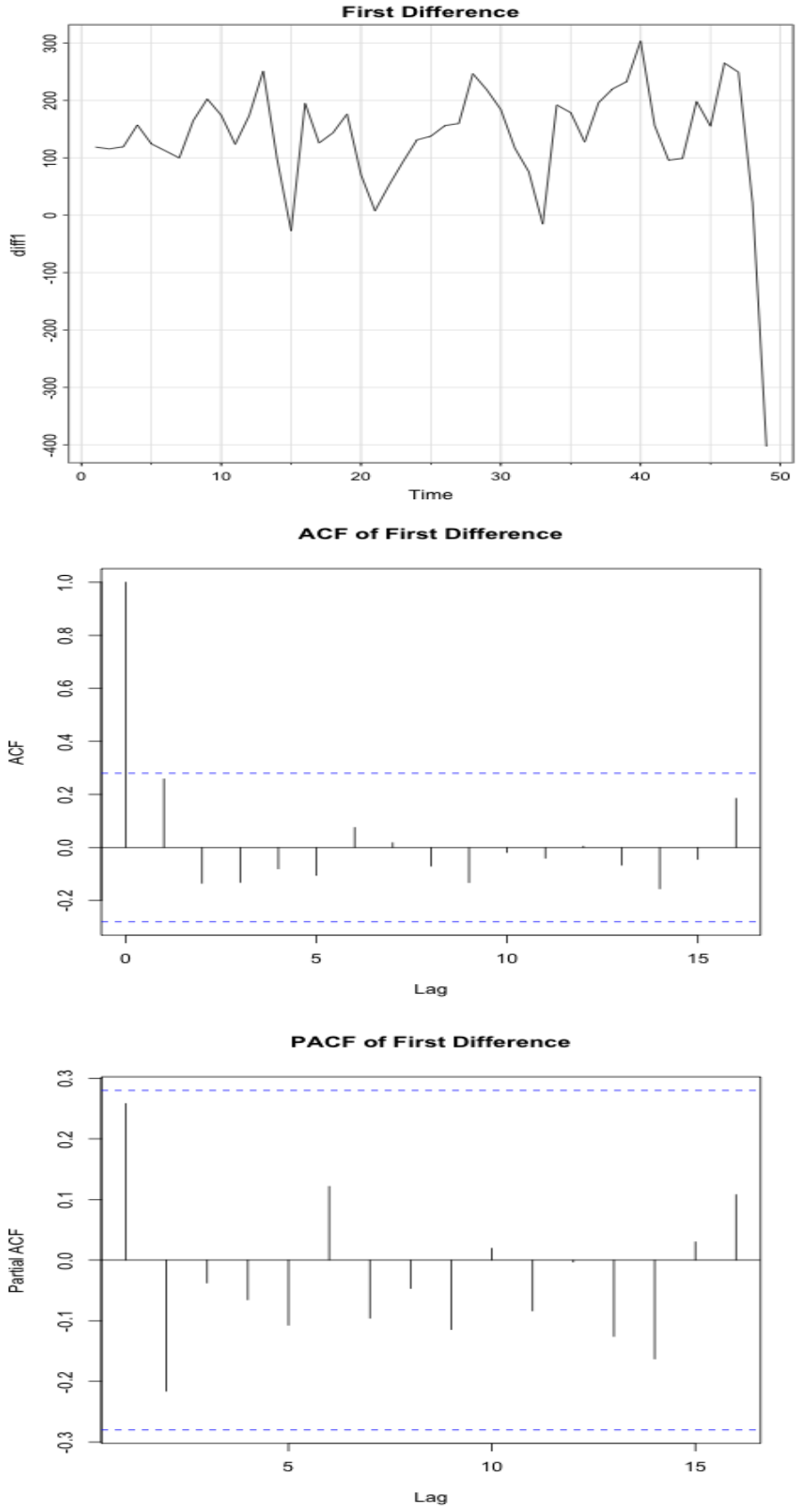

Figure 16: Plot of time series (eu15), ACF, and PACF with one differencing.

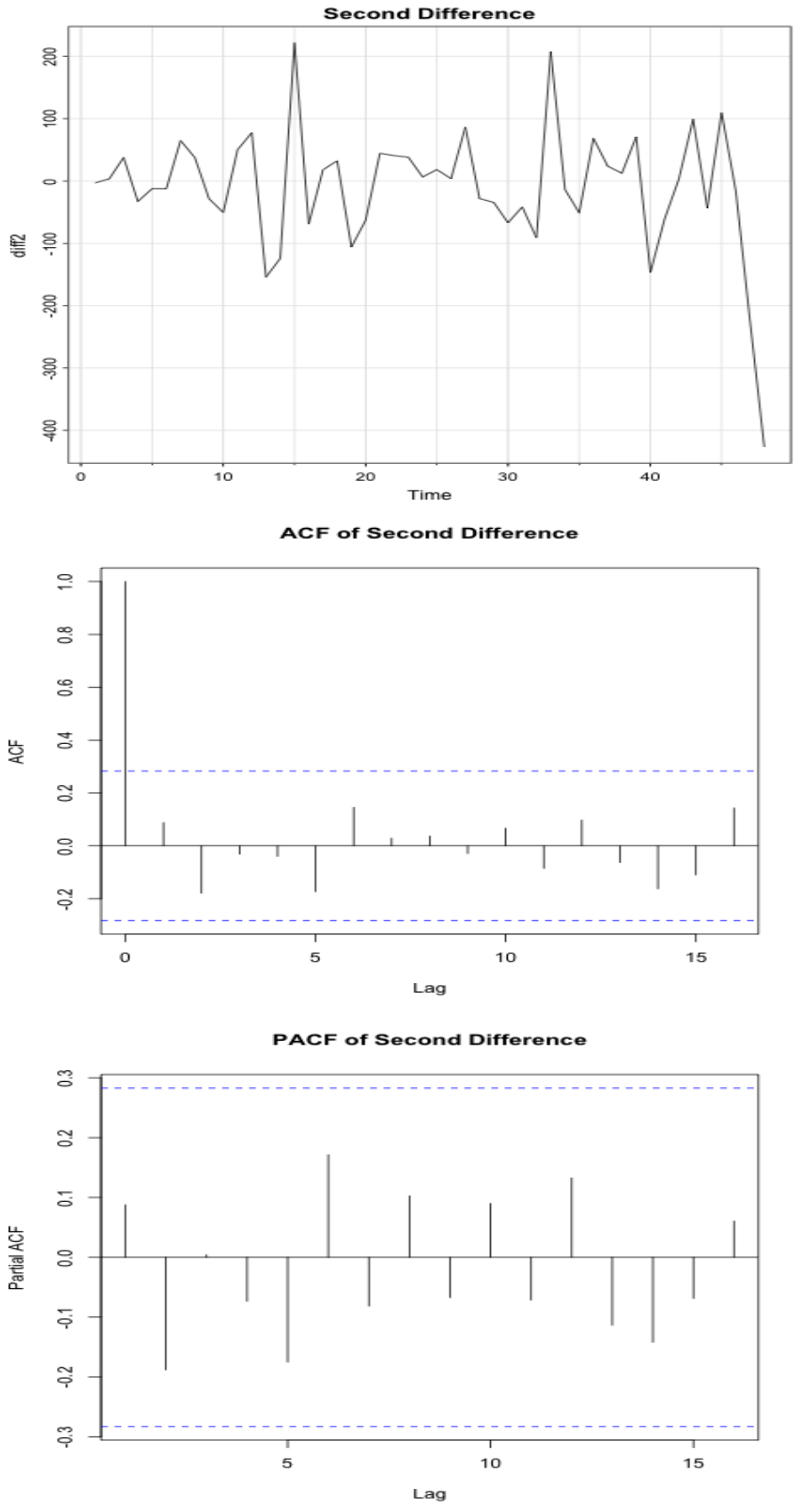

Figure 17: Plot of time series (eu15), ACF, and PACF with two differencing.

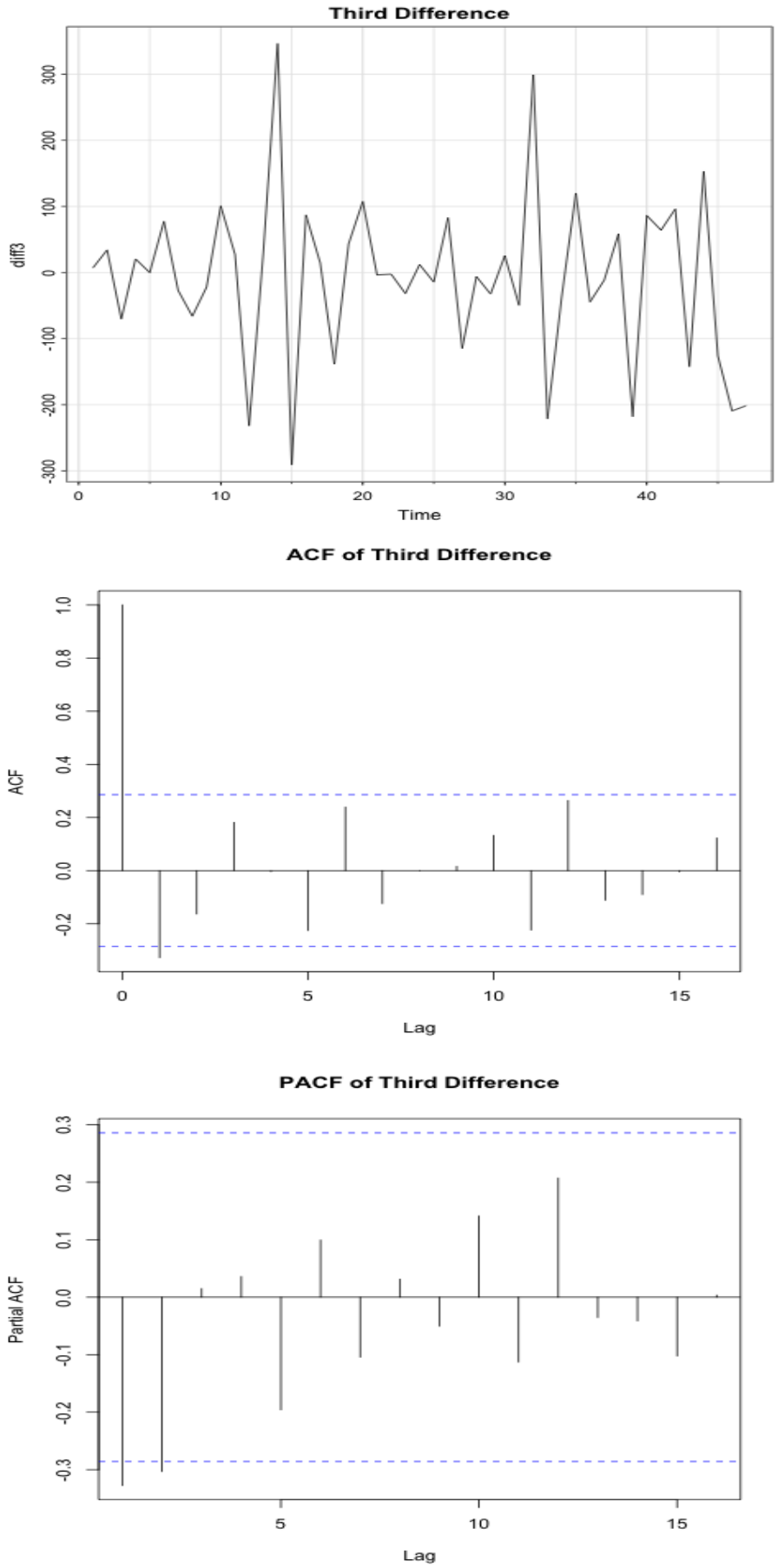

Figure 18: Plot of time series (eu15), ACF, and PACF with three differencing.

0.0.3 Question 3 - Investigating Different ARMA models

Model	df	AIC	BIC
ma1	3	589.2094	594.8848
ma2	4	589.1047	596.6720
ma3	5	588.4865	597.9456
ar1	3	592.7084	598.3838
ar2	4	587.8486	595.4159
ar3	5	589.8126	599.2717
arma11	4	590.1944	597.7616
arma21	5	589.7361	599.1952
arma31	6	591.8485	603.1994
arma12	5	589.7361	599.1952
arma13	6	591.8485	603.1994

Table 10: AIC and BIC Values for Various ARIMA Models for the original time series with one differencing

Model	df	AIC	BIC
ma1	3	542.9974	548.6110
ma2	4	544.6056	552.0904
ma3	5	546.4000	555.7560
ar1	3	545.5871	551.2007
ar2	4	545.3718	552.8566
ar3	5	547.1611	556.5171
arma11	4	544.6793	552.1641
arma21	5	546.5174	555.8734
arma31	6	548.5026	559.7298
arma12	5	546.5174	555.8734
arma13	6	548.5026	559.7298

Table 11: AIC and BIC Values for Various ARIMA Models with outlier removed for one differencing.

0.0.4 Question 3 - Investigating Different Forecasts using our Model

Here we show the time series forecast without the outlier fig. 19.

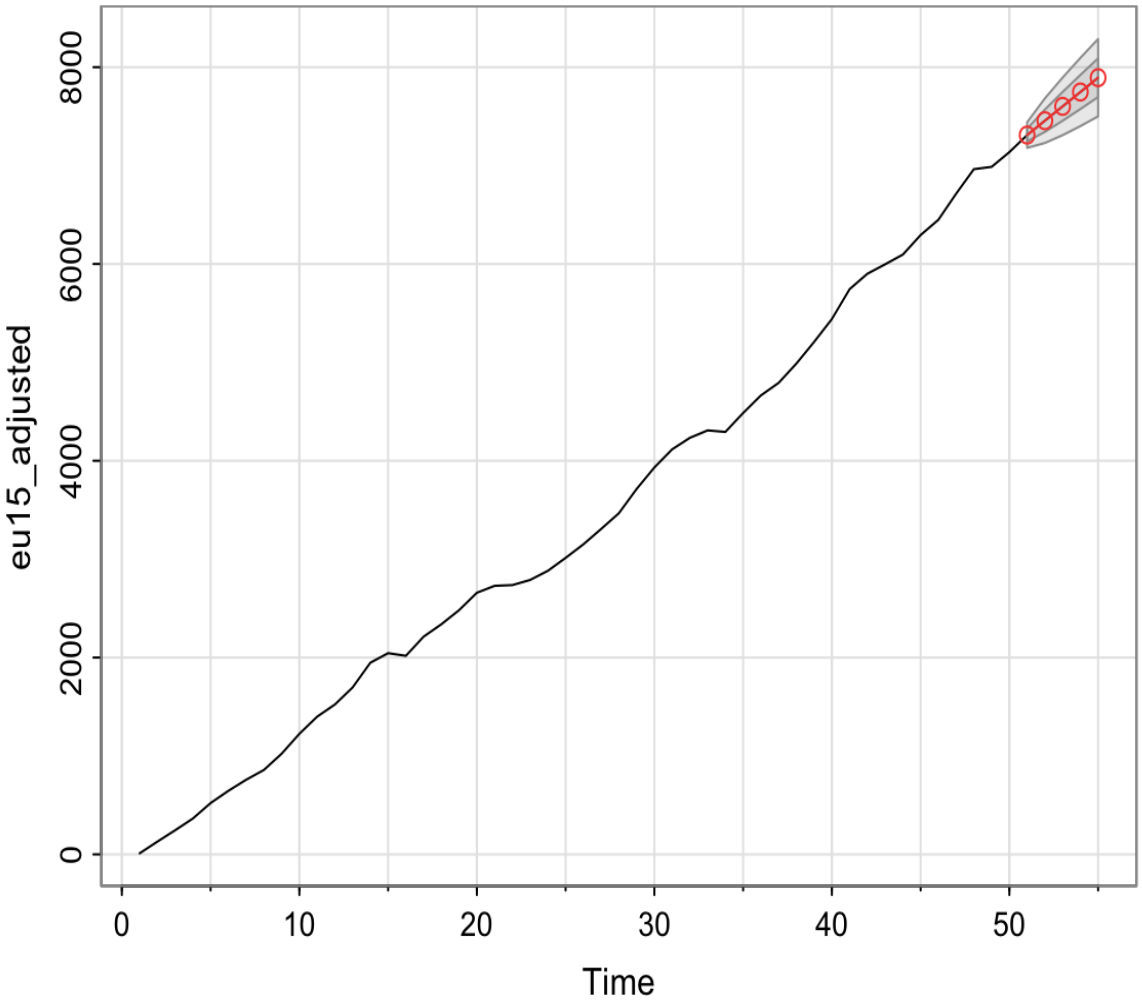

Figure 19: Time series forecast without the outlier

0.0.5 Question 4 - Investigating co-integration

> adf.test(residuals)

Augmented Dickey-Fuller Test

data: residuals Dickey-Fuller = -1.7711, Lag order = 4, p-value = 0.6713 alternative hypothesis: stationary

using PO test we get value of 0.01 meaning they are cointegrated

> po.test(cbind(log(log_income),log(log_consumption)))

Phillips-Ouliaris Cointegration Test

data: cbind(log(log_income), log(log_consumption)) Phillips-Ouliaris demeaned = -31.531, Truncation lag parameter = 0, p-value = 0.01

if we use johansen test we get eigen vectors which create the same timesseries as our residuals from ols. meaning by adf there is no correlation!!!!

> summary(johansen_test)

Johansen-Procedure

Test type: maximal eigenvalue statistic (lambda max) , with linear trend

Eigenvalues (lambda): [1] 0.1625319555 0.0005359794

Values of test statistic and critical values of test:

test 10pct 5pct 1pct r <= 1 | 0.05 6.50 8.18 11.65 r = 0 | 17.21 12.91 14.90 19.19

Eigenvectors, normalised to first column: (These are the cointegration relations)

log_income.l2 log_consumption.l2 log_income.l2 1.0000000 1.0000000 log_consumption.l2 -0.9851151 0.6333639

Weights W: (This is the loading matrix)

log_income.l2 log_consumption.l2 log_income.d -0.3379134 6.009201e-05 log_consumption.d -0.1188643 -9.087631e-04

If you made it here it's time for a pint