Analysing average income per person in Ireland with respect to year from 1950 to 2019

Jack Curtin
16181484

Time Series Project
May 4, 2021

Chapter 1

Introduction

In this report we are going to investigate a time series dataset and comment on our findings. The dataset which we will be examining was downloaded from `https://wid.world/data/` and it describes the average income per person in Ireland from 1950 to 2019. This dataset contains two columns where the first column indicates the year and the second column indicates the average income per person (in €) for that given year. The aim of this report is to create a model for predicting the average income per person in Ireland with respect to year and we will do so using "R software" (R). We begin by removing the last 10% of the dataset so we can use the remaining 90% as a basis for creating a predictive model for future values, we call this new dataset `dataTSNew`. Our first step is to check if our data can be rationally modelled, which is to say that the data meets certain criteria which we will explore in the coming section.

1.1 Checking the Data

It is essential that we ensure the data we are looking at meets the requirements of a time series which can be modelled. We will begin by looking at a plot of our data (which has been converted to a time series in R) to see if there are any immediate visual observations. From **Figure 1.1** we can see a clear upward trend in our data, which is to be expected of course, with the exception of some years from 2007 where we see a temporary downward trend (a direct consequence of the unprecedented *housing market crash*), although we see the upward trend appears to have corrected itself from around 2013.

The first requirement we will check is that our data is not white noise, to do so we plot the *auto correlation function* (ACF) of the data to see if we have any significant correlations, this can be seen in **Figure 1.2**. What we notice from this plot is that we have a number of significant correlations which implies our data is not white noise. The second requirement is that the data

is not a random walk, to check this we plot the ACF for the *differenced data* and we notice a number of significant differenced correlations (from the plot in **Figure 1.3**) which implies our data is not a random walk. As these two requirements are satisfied we can now look into checking for stationarity and transforming our data to a stationary time series if necessary.

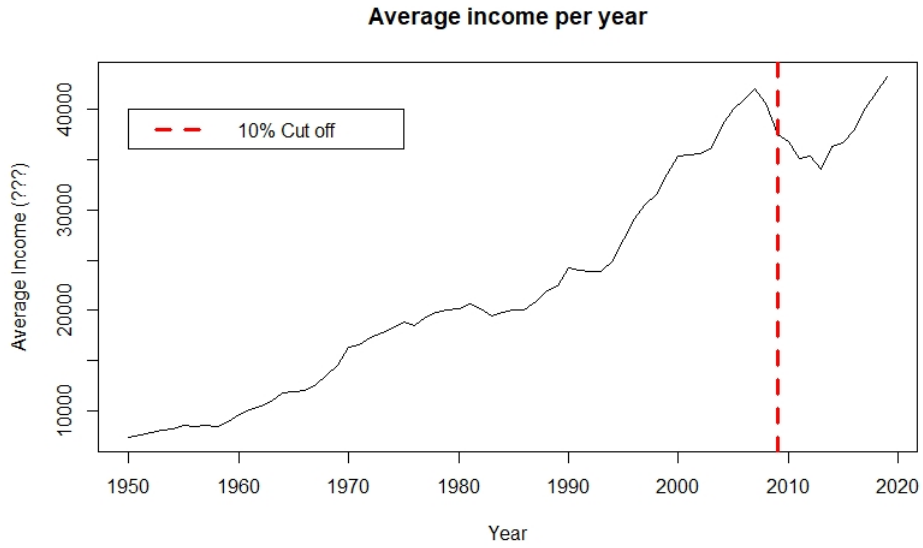

Figure 1.1: Plot of original time series data with the 10% cut off point from 2009 indicated.

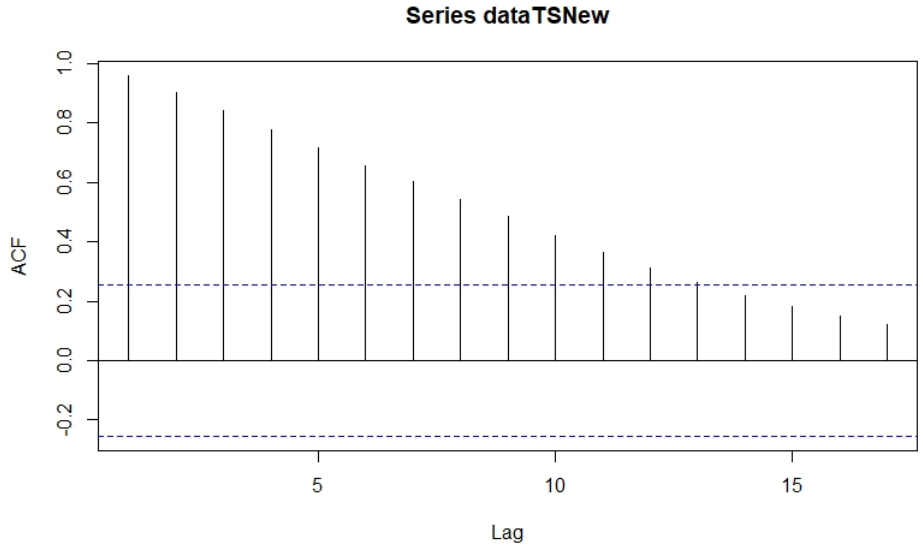

Figure 1.2: ACF of the data.

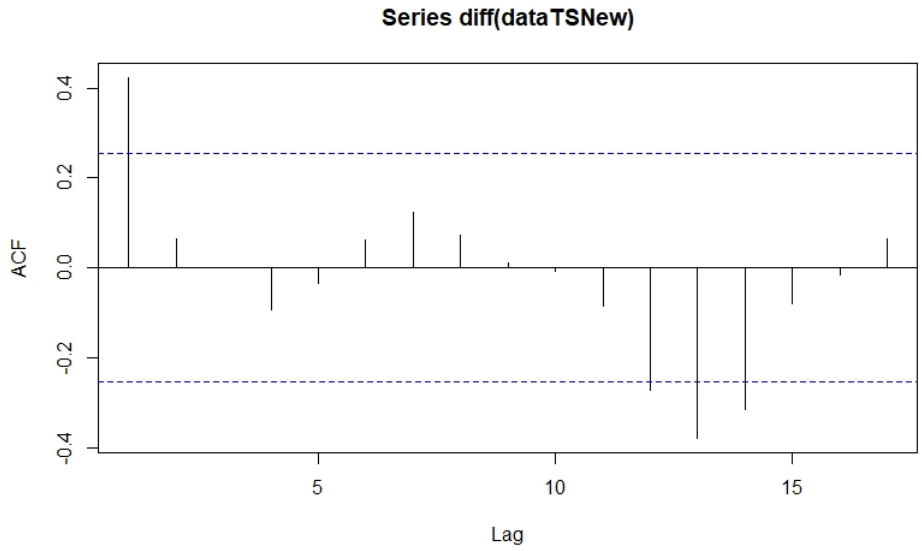

Figure 1.3: ACF of the differenced data.

Chapter 2

Analysing the Data

We will now look into analysing our data to see if it is stationary or if we need to transform our data to make it stationary. We need our data to be stationary to allow us to make inferences about its structure. We notice from **Figure 1.1** that our data appears to have a positive linear trend and we can confirm this hypothesis by looking at the *Augmented Dickey-Fuller Test* (ADF test). The ADF test for our data gives us a p-value of 0.3961, which is much larger than 0.05 and therefore confirms our hypothesis that our data is not stationary. Thus we will begin the process of transforming our data to make it stationary so that we can create some predictive models.

2.1 Transforming the Data

To transform our data to a stationary time series we will begin by interpreting the *Box-Cox power transformation*. This transformation is useful for making a time series more normally distributed, and we can test it using the R code below which outputs **Figure 2.1**.

```
BC <- BoxCox.ar(dataTSNew, lambda=seq(-2,2,0.1))
BC$mle # MLE is lambda = -0.1
BC$ci # [-0.5, 0.2]
```

We see from the R code output that our maximum likelihood estimate value is $\lambda = -0.1$, although we will instead use the value $\lambda = 0$ as it is included in our confidence interval $[-0.5, 0.2]$ and is a more idealistic value to incorporate than -0.1. Choosing the value $\lambda = 0$ implies that we can log transform our time series and doing so we find that our data is in fact still not stationary as our ADF value for the log transformed data set, which we call `dataTSLog`, is 0.7013 which is far greater than 0.05. Therefore we will difference our new log transformed data set, which we

call `dataTSLogDiff` and we find it's corresponding ADF value to be 0.07713 which is still greater than 0.05 although not by a large margin. Therefore we will difference this once more, calling it `dataTSLogDiff2` and when we examine **Figure 2.2**, which is a plot of `dataTSLogDiff2`, we can see that our data seems to be stationary. This assumption is confirmed as we find the ADF value of `dataTSLogDiff2` to be much smaller than 0.05, therefore confirming that our data is stationary.

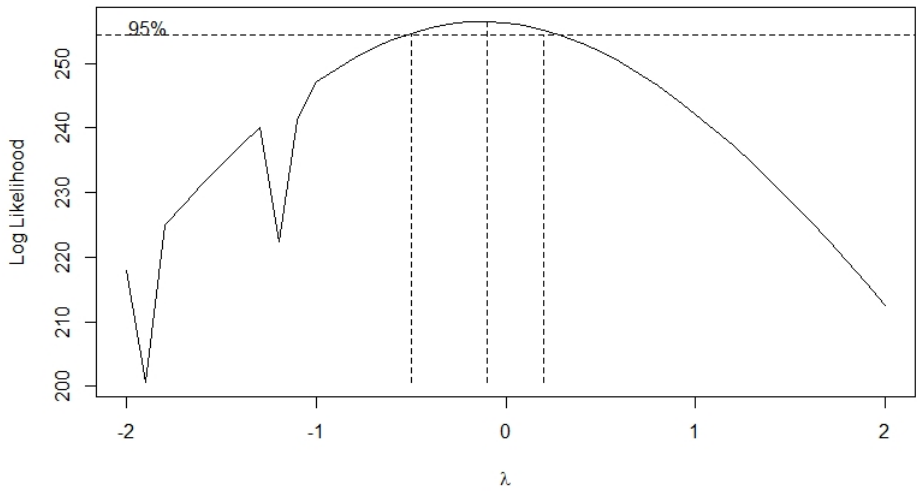

Figure 2.1: Box-Cox transformation of the data.

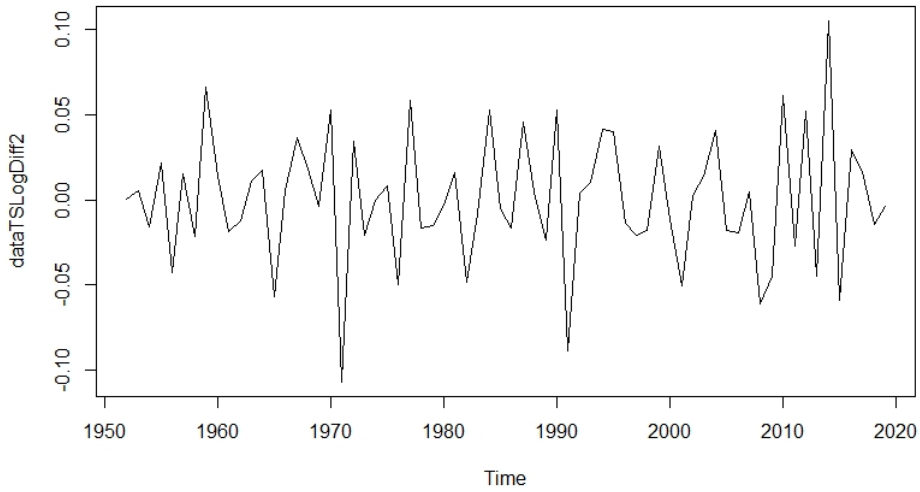

Figure 2.2: Plot of `dataTSLogDiff2`

Chapter 3

Modelling the Data

We can use our stationary time series `dataTSLogDiff2` to create predictive models. Firstly we'll examine the ACF of our time series (**Figure 3.1**) which we can use to determine if our model has any MA(q) values. Our ACF only one significant auto correlation located at lag-1 which implies that each year is strongly correlated with the previous year and from this ACF plot there doesn't appear to be any seasonality in effect (such as a significant auto correlation at lag-12). Based on this rationality, we can assume our predictive model will not be affected by seasonality and will be in the form of an ARIMA. We can also examine the *Partial ACF* (PACF) which is show in **Figure 3.2** and it is the residual correlation between two random variables after removing the effects of other variables. We can use the PACF to determine if our model has any AR(p) values. What we notice is a significant auto correlation located at lag-1 and lag-14. Factoring in both our ACF and PACF results we can reasonably assume that our model will likely be in the form of an ARIMA$(p, 2, q)$, where p and q need to be determined and we choose $I = 2$ as we had to difference our time series twice to impose stationarity.

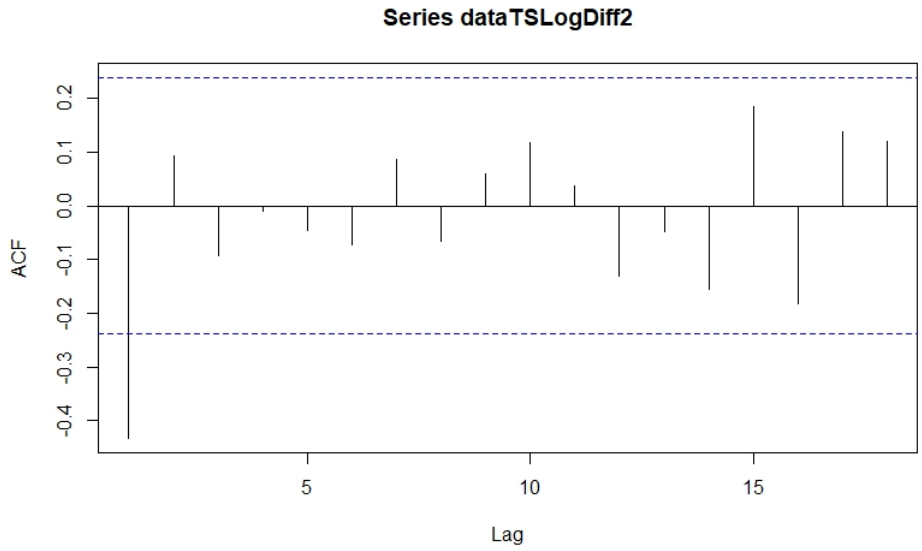

Figure 3.1: ACF of `dataTSLogDiff2`

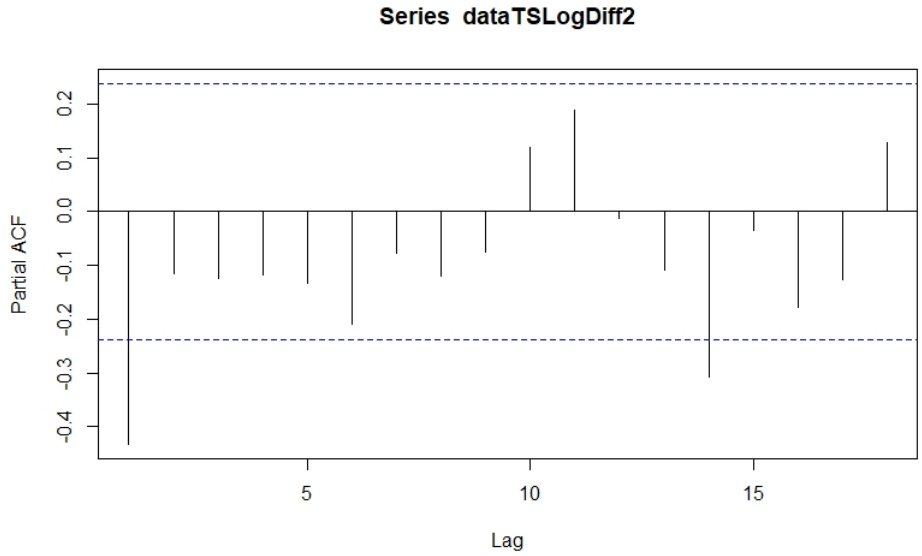

Figure 3.2: PACF of `dataTSLogDiff2`

3.1 Choosing Models

To determine which model to choose we can interpret an *Extended ACF* (EACF) using the `eacf()` function in R. An EACF is used for determining the orders p and q of an ARMA process. The EACF R code and corresponding output is shown below and it gives us suggestions as to

which values to choose based on the positions marked with an `o`.

```
eacf(dataTSLogDiff2, ar.max=8, ma.max=8)

AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x * o o o o o o o o o  o  o  o
1 x x o o o o o o o o o  o  o  o
2 x * o o o o o o o o o  o  o  o
3 x x o o o o o o o o o  o  o  o
4 x * o o o o o o o o o  o  o  o
5 x o o o o o o o o o o  o  o  o
6 x x o x o o o o o o o  o  o  o
7 x x o o o o o o o o o  o  o  o
```

Based on the output, we can see the majority of suggestions are of the form ARMA$(p, 1)$ and as such we've marked the more relevant suggestions with a `*`. We can also use the `armasubsets()` function in R, which shows the top eight models ranked based on the *Bayesian Information Criterion* (BIC), which uses the formula

$$\text{BIC} = -2l(\hat{\phi}, \hat{\theta}) + m\ln(n),$$

which outputs **Figure 3.3** below and provides a measure of fit with a penalty for complexity. Ideally we wish to use a model with a low BIC value, and from our output we can see MA(1) has the lowest value followed by ARMA(1,1) and ARMA(1,8) (with significant e_{t-1} and e_{t-8} terms). Taking this output into consideration and factoring in our EACF output above, we will look at the following three models and interpret which of them is best suited to our data:

- ARIMA$(0, 2, 1)$
- ARIMA$(2, 2, 1)$
- ARIMA$(4, 2, 1)$

We will determine which model is best by factoring in the *Akaike Information Criterion* (AIC), which is similar to the BIC from above and which has the following equation

$$\text{AIC} = -2l(\hat{\phi}, \hat{\theta}) + 2m,$$

it also provides a measure of fit with a penalty for complexity where a lower AIC value is generally considered to be better.

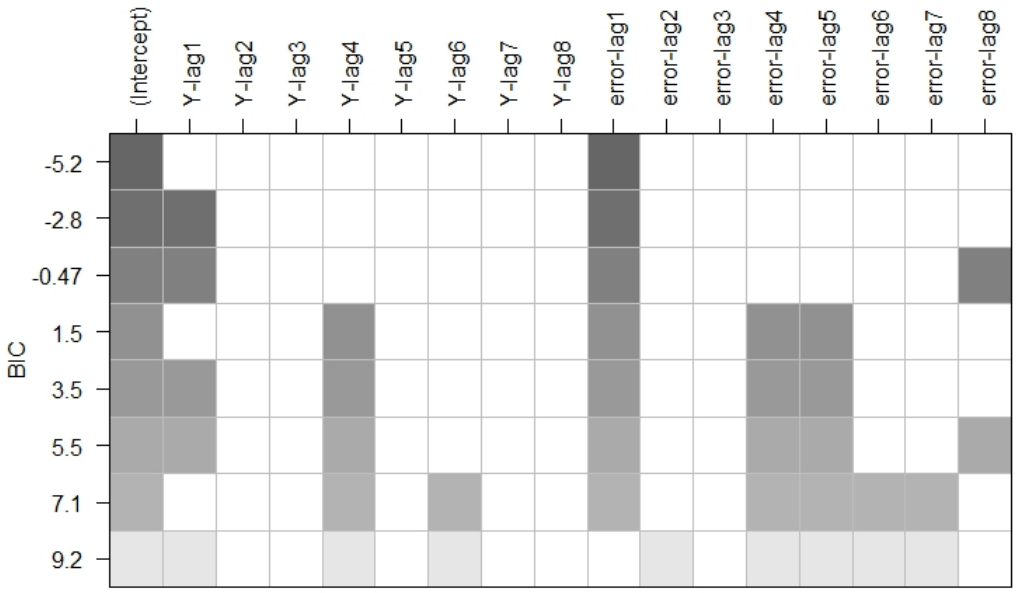

Figure 3.3: `armasubsets(dataTSLogDiff2)`

3.1.1 Model 1 $\rightarrow$ ARIMA$(0, 2, 1)$

The first model we will consider is an ARIMA$(0, 2, 1)$, which is also know as a IMA$(2, 1)$. We find the AIC value for this model to be -230.02, which is important to note when we compare this model to other models. We can see that the residuals from this model appear to be normal as evident from a histogram and a QQ-Plot of our residuals shown in **Figure 3.4**. We can also test for normality using the *Shapiro-Wilk* test , which has a value of 0.503 for this model, so we accept the null hypothesis that the data is normal. Looking at a plot of the fitted residuals we can see that the points seem randomly scattered with no evidence of patterns or non-constant variance. The ACF plot of the residuals does seem to have some significant values at lags 12, 13 and 14 and when we look at the p-Values from the *Ljung-Box* test we see that some of them are below the 0.05 reference line which suggest that some residuals are not white noise. Each of these three plots are seen in **Figure 3.5**. Therefore we conclude that perhaps this model wouldn't be the best choice.

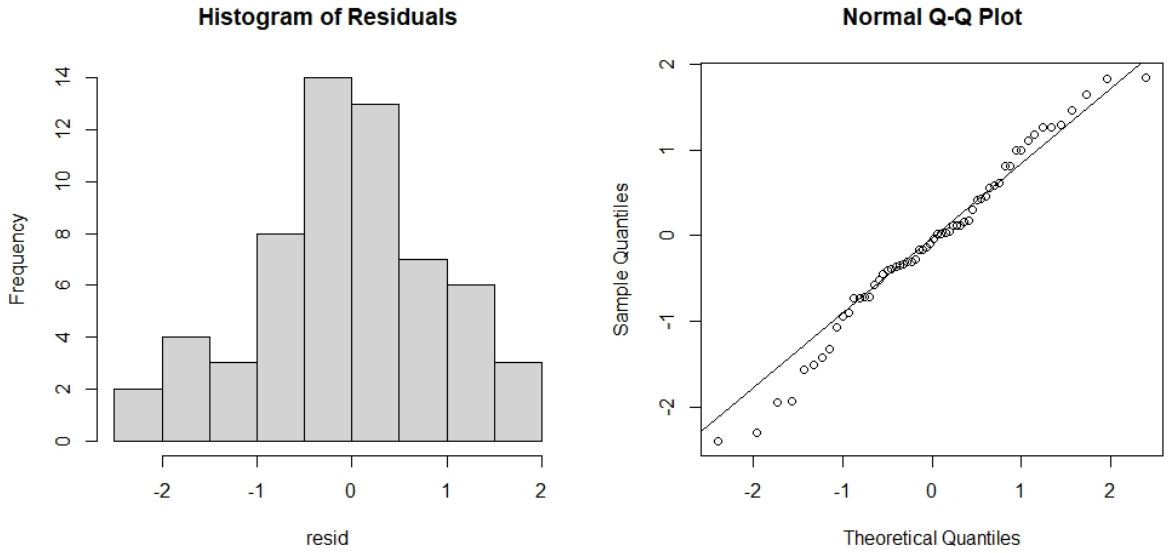

Figure 3.4: Histogram and QQ-Plot of our residuals for Model 1.

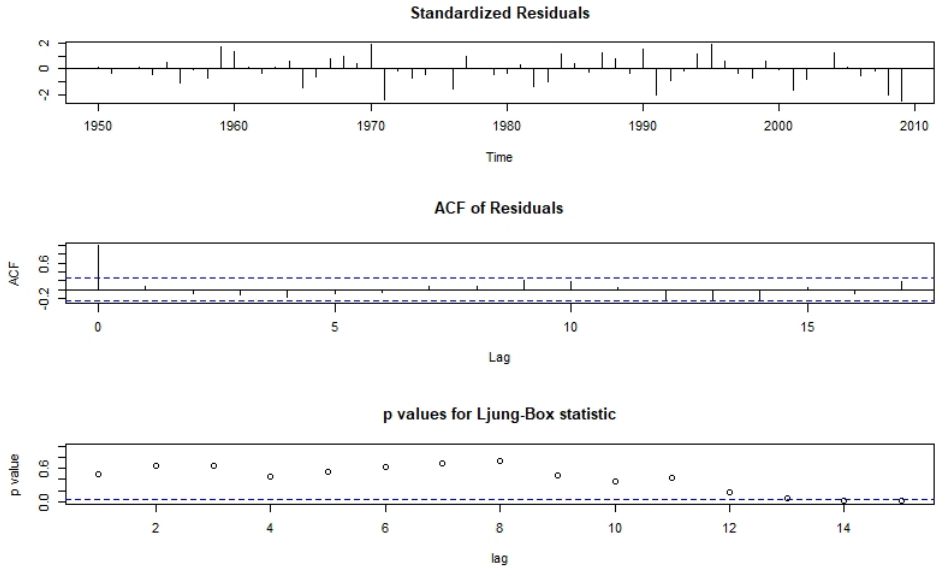

Figure 3.5: Time Plot, ACF and Ljung-Box p-values for Model 1.

3.1.2 Model 2 $\rightarrow$ ARIMA$(2, 2, 1)$

The second model we will consider is an ARIMA$(2, 2, 1)$. We find the AIC value for this model to be -230.71, which is marginally better than the AIC value for our IMA$(2, 1$ model. We can see that the residuals from this model appear to be normal as evident from a histogram and a

QQ-Plot of our residuals shown in **Figure 3.6**. The Shapiro-Wilk test gives us a value of 0.4176 for this model, so we accept the null hypothesis that the data is normal. Looking at a plot of the fitted residuals we can see that the points seem randomly scattered with no evidence of patterns or non-constant variance. The ACF plot of the residuals does seem to have one significant value at lag 14, although when we look at the p-Values from the Ljung-Box test we see that all of the p-values are well above the 0.05 reference line which supports the hypothesis that the residuals are white noise. Each of these three plots are seen in **Figure 3.7**. Therefore we conclude that this model would be a suitable choice.

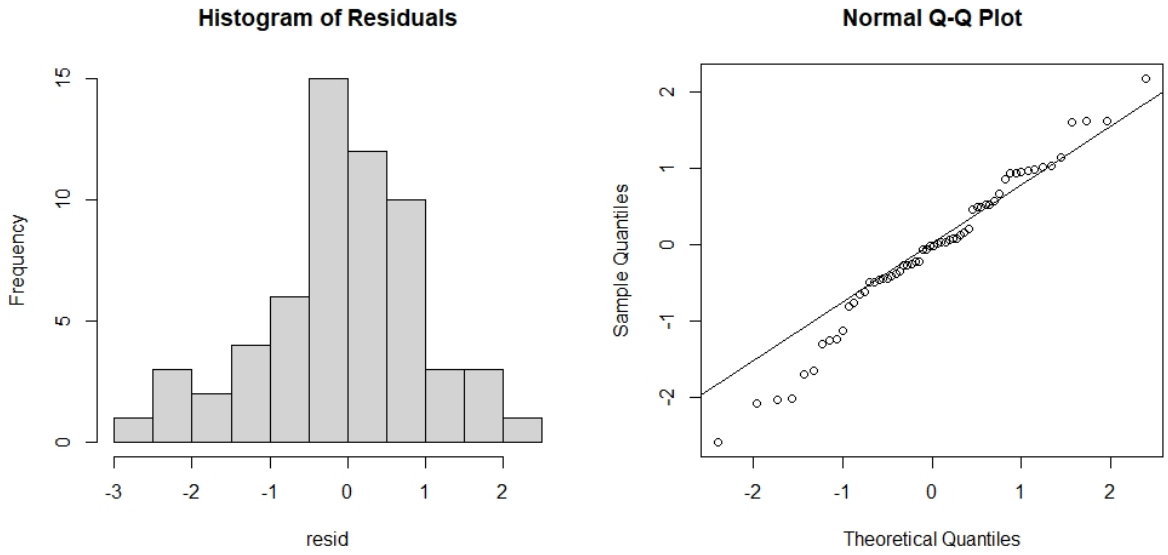

Figure 3.6: Histogram and QQ-Plot of our residuals for Model 2.

3.1.3 Model 3 $\rightarrow$ ARIMA$(4, 2, 1)$

The third model we will consider is an ARIMA$(4, 2, 1)$, which is also know as a ARI$(2, 2)$. We find the AIC value for this model to be -226.93, which is not as good as our IMA$(2, 1)$ or ARIMA$(2, 2, 1)$ AIC values. We can see that the residuals from this model appear to be normal (all values are within [-3,3]) as evident from a histogram and a QQ-Plot of our residuals shown in **Figure 3.8**. The Shapiro-Wilk test gives us a value of 0.0.6229 for this model, so we accept the null hypothesis that the data is normal. Looking at a plot of the fitted residuals we can see that the points seem randomly scattered with no evidence of patterns or non-constant variance. The ACF plot of the residuals does seem to have one significant value near lag 14, which is similar to our previous model, although when we look at the p-Values from the Ljung-Box test we see that all of the p-values are well above the 0.05 reference line which supports the hypothesis that the residuals are white noise. Therefore we conclude that this model would also be a suitable choice.

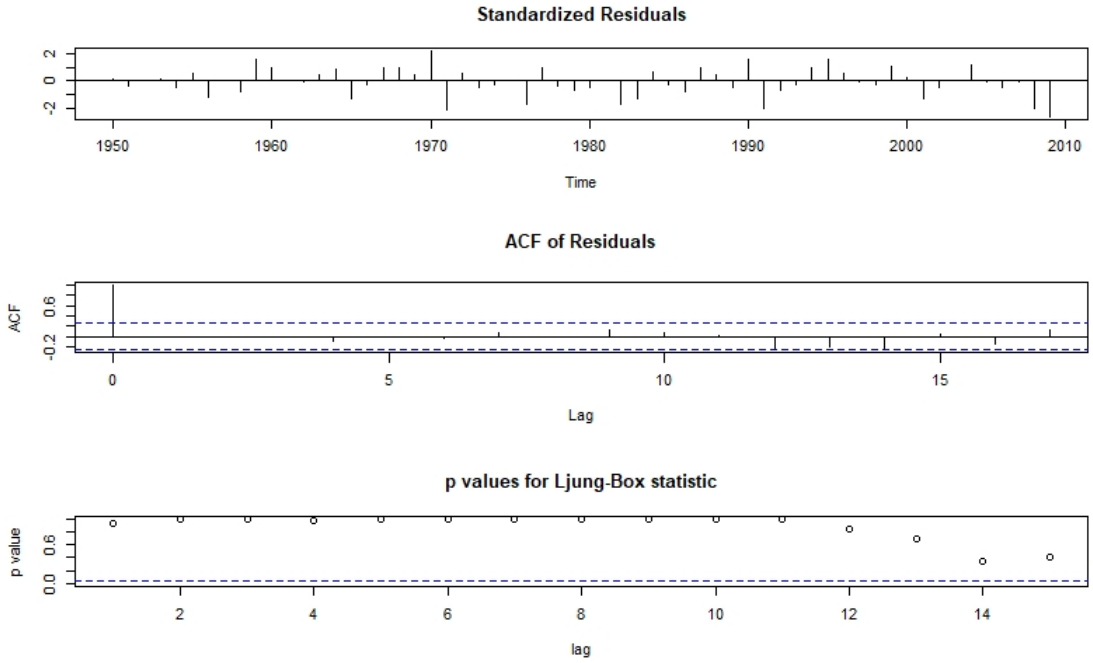

Figure 3.7: Time Plot, ACF and Ljung-Box p-values for Model 2.

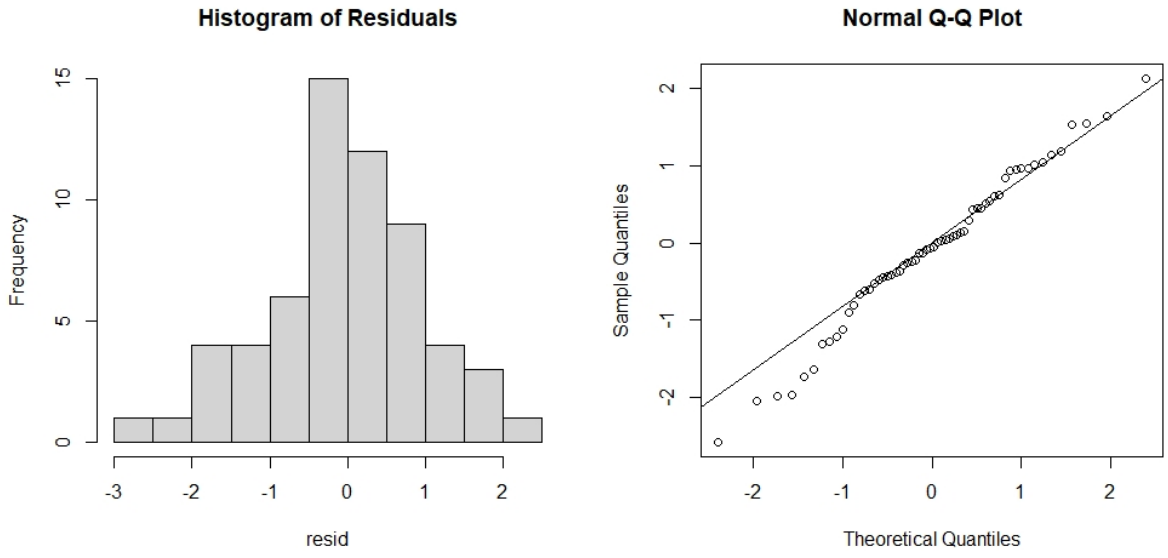

Figure 3.8: Histogram and QQ-Plot of our residuals for Model 3.

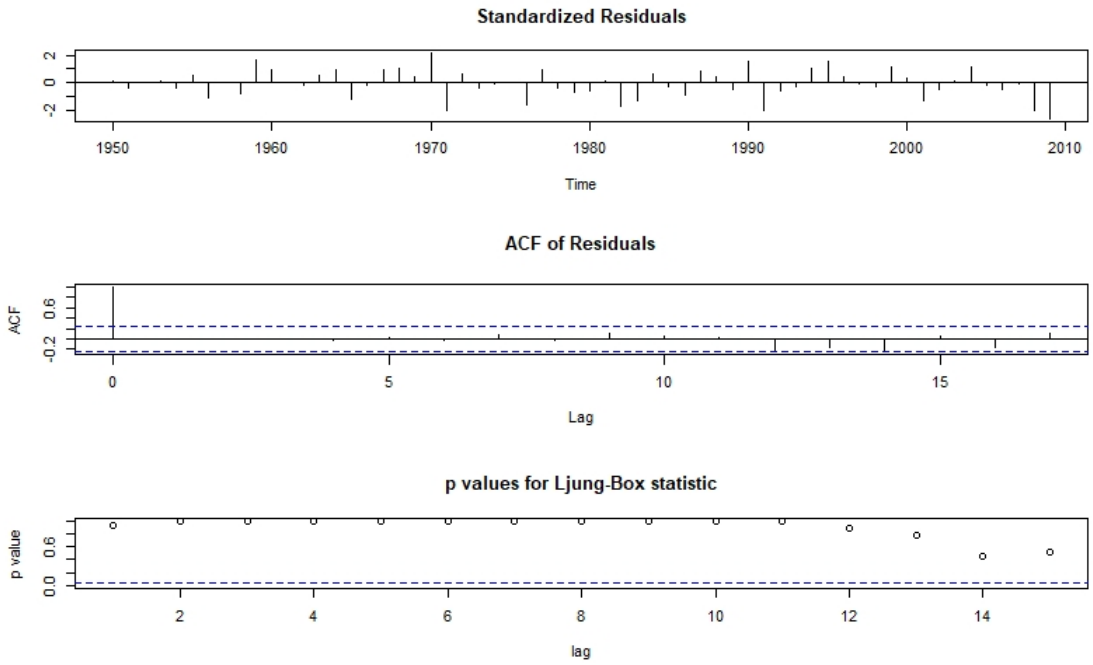

Figure 3.9: Time Plot, ACF and Ljung-Box p-values for Model 3.

Chapter 4

Conclusion

We therefore conclude that the model which best suits our data is Model 2 $\rightarrow$ ARIMA$(2, 2, 1)$. We came to the conclusion that this model fits best after taking out the last 10% of our original dataset and turning the remaining 90% into a time series. Next, after ensuring our data was not a random walk or white noise, we began the process of making our time series stationary which began with a Box-Cox transformation that suggested we log-transform our time series. Using this log-transformed time series as a basis, we found that differencing this time series twice made our data stationary as it gave us an ADF value far smaller than 0.05. We then used an EACF to consider a number of ARMA models which fitted our data and noted each model's AIC and BIC values and compared them with each other amongst other characteristics, thus leading us to determine that Model 2 $\rightarrow$ ARIMA$(2, 2, 1)$ best fits our model. We find the relevant coefficients of our model to be the following:

```
Coefficients:
          ar1      ar2     ma1
      0.4447   0.0333  -1.00
s.e.  0.1410   0.1456   0.16

sigma^2 estimated as 0.0009382:  log likelihood = 118.35,
aic = -230.71
```

We can see a plot of our chosen model in **Figure 4.1** which shows our original values plotted in green and our predicted values plotted in black. The red lines on the plot describe the confidence band of our predicted values, and as we can see the original values are well within this confidence band. **Figure 4.2** shows a close-up of our plot to make it easier to distinguish between our predicted and actual values. We can see that whilst our prediction isn't necessarily exactly the

same as our original values; it is reasonably close and it also captures the general upward trend of our original values.

An argument to support our predicted values straying slightly from the original values would be the housing market crash (seen from $2007 - 2008$) that we mentioned previously. Essentially this housing market crash is one of the significant reasons as to why the general upward trend for average income peaked in 2007 and began an unusual downward trend until around 2013. Our predicted and original values do however seem to converge for the later years as the upward trend corrects itself, suggesting that our predicted and actual values are likely to become reasonably indistinguishable in the future, unless of course we are struck with another unprecedented economic disaster.

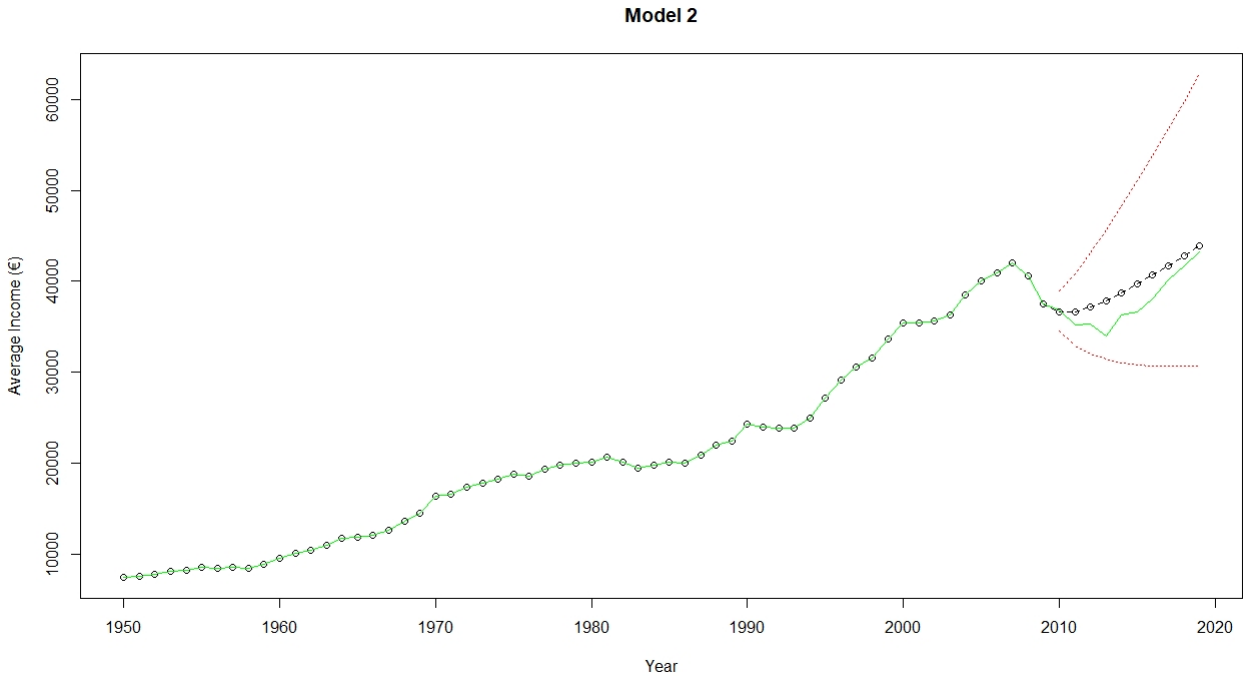

Figure 4.1: Plot of our predictive model ARIMA$(2, 2, 1)$ (black) against our actual values (green).

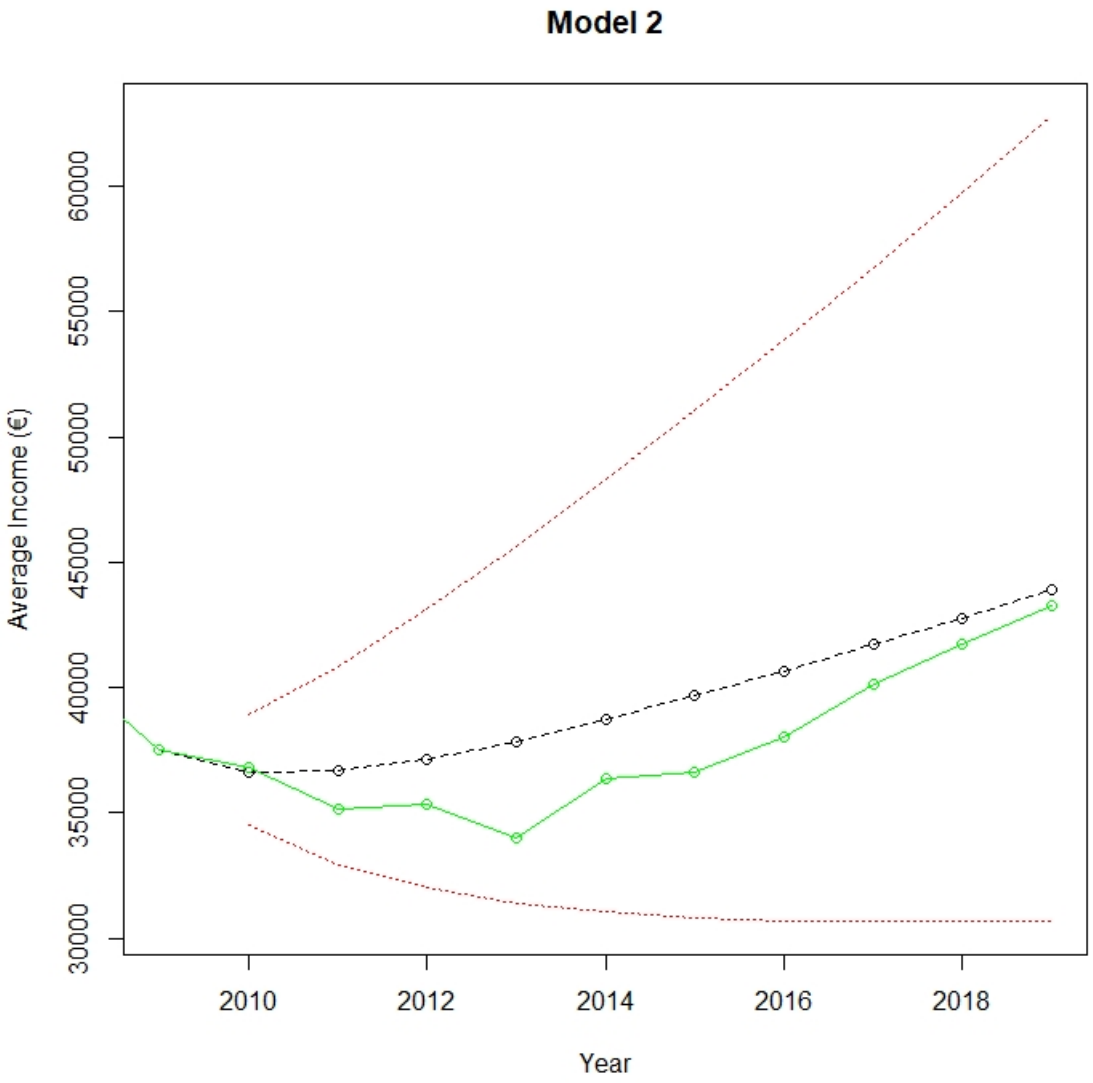

Figure 4.2: Close-up plot of our predictive model ARIMA$(2, 2, 1)$ (black) against our actual values (green).