Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

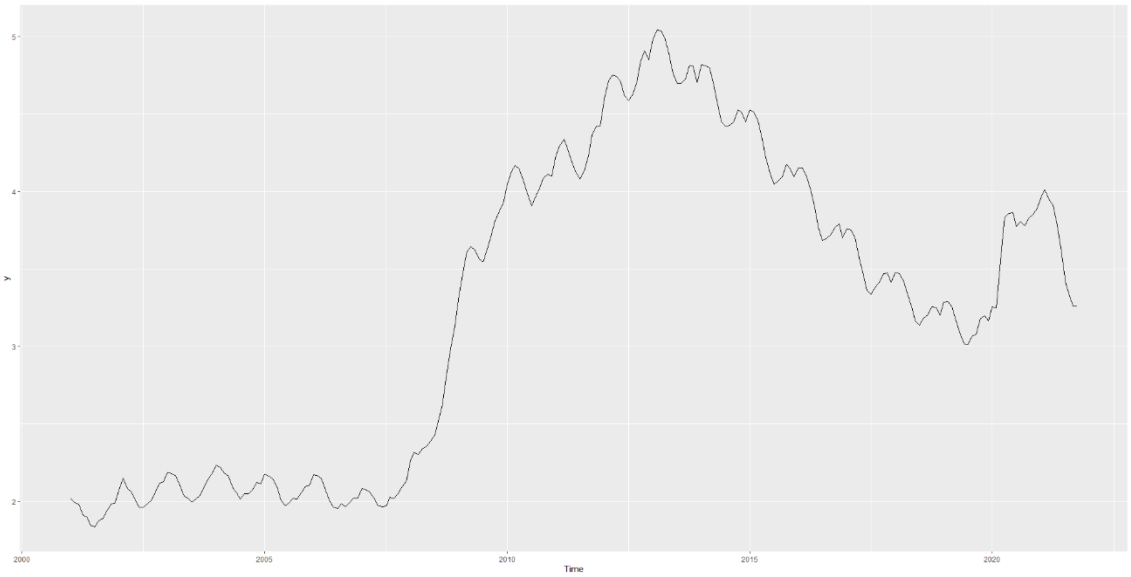
*Assignment 2:*
*Forecasting*

## Introduction

In this report we will try to forecast the total number of unemployed people in Spain for the month of November.

This report has been created due to the strange situation the Spanish unemployment is in right now. The 2008 global financial crisis hit Spain especially hard. Most developed countries recovered years ago from the 08 crisis, while Spain was still struggling with debt payments and record high unemployment numbers. When things started to look better for Spain, as unemployment numbers have consistently decreased since 2014, the covid pandemic hit, causing mass unemployment. In recent months, the economy has started to recover, and we have seen a steep decrease in unemployment, although absolute numbers are still well above pre-crisis and pre-pandemic numbers.

To forecast the total unemployment in Spain, we will use two different models. Firstly, we will use a Seasonal ARIMA model. Spain's heavy reliance on the tourism sector makes the unemployment numbers very cyclical and season. Every summer, unemployment drops, but increases after the summer has ended, therefore we need to introduce the seasonal element in our ARIMA model.




The second model we will use is a dynamic regression model with an intervention variable. We use an intervention variable as we try to accurately model the effect the pandemic has on unemployment numbers.

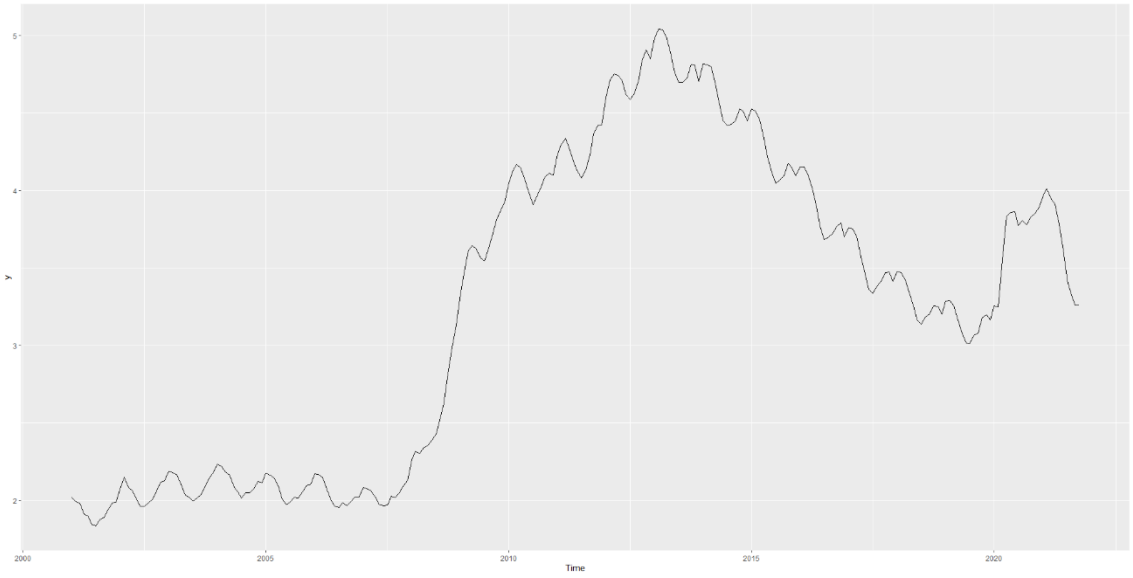
Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

# SARIMA model

Several steps need to be taken in order to identify the best SARIMA model to forecast the total unemployment rate.

Firstly, we plot the data to visually analyse whether the time series is stationary or non-stationary.



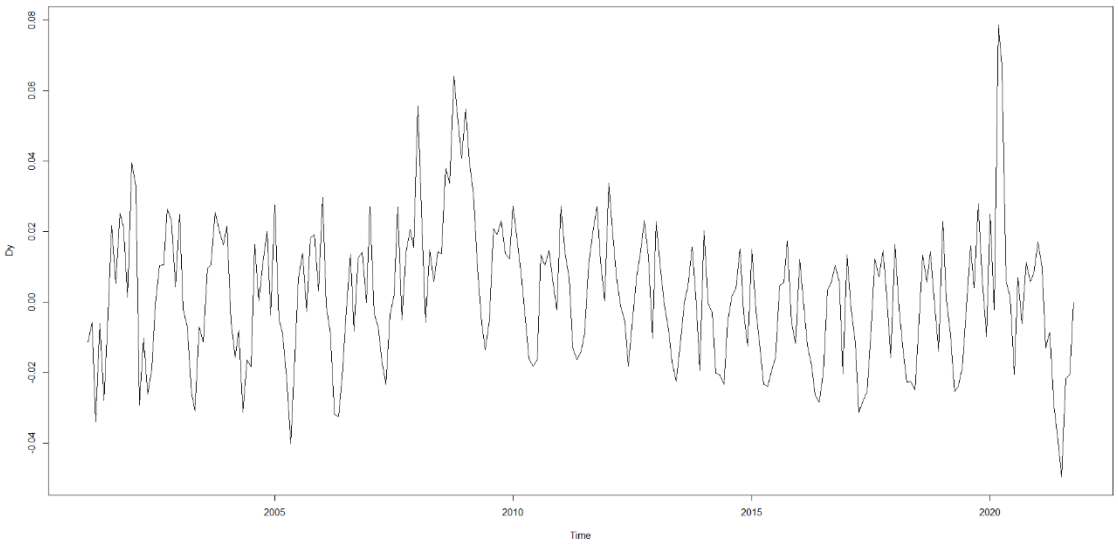

From the above graph, we can see that from around 2013 onwards, there has been an element of seasonality, as the unemployment rate lowers every summer, and then increases afterwards, therefore we can hypothesize that the time-series is non-stationary, as the unemployment rate is dependent on the time at which the series is observed.

We can also use a statistical check to test our hypothesis. We will use an augmented Dickey-Fuller test, where the null hypothesis is that the ts is non-stationary, therefore if we fail to reject $H_0$ (p-value > 0.05) we can conclude that the ts is non-stationary. After using the D-F test, we observe that we fail to reject $H_0$, therefore we conclude that the ts is non-stationary.

```
data:  y
Dickey-Fuller = -0.72766, Lag order = 6, p-value = 0.9672
alternative hypothesis: stationary
```

There are several ways to deal with a non-stationary ts, but we will transform the data by differentiating the data at least once.

Once we differentiate the data once, we can observe that the data is now stationary, that is, the unemployment rate is no longer dependent on the time at which the series is observed.

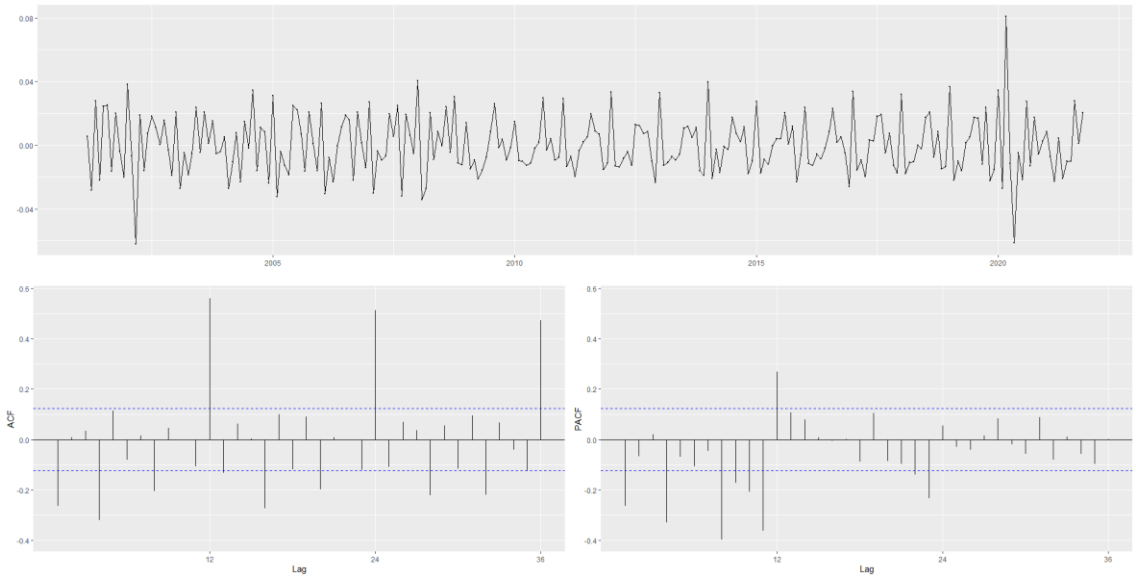



When using the ARIMA model, we will have to include the d parameter, to indicate that we want the ts to have one differentiation, making the ts stationary. We can also use the D-F test to statistically prove we no longer have a non-stationary ts.

```
data:  Dy
Dickey-Fuller = -20.463, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

We now divide the data set into training and test, taking the last year for test, while the rest of the dataset will be for training.

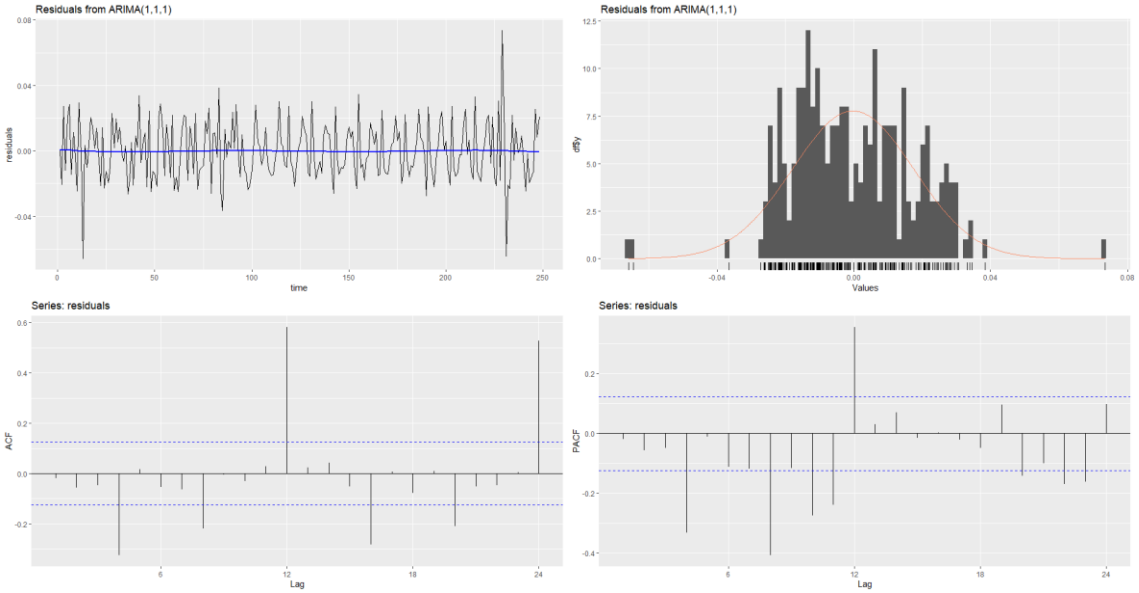
Now that we have a stationary ts, we can proceed to analyse the ACF and PACF.




These are the results when we want to analyse the ACF and PACF. There is a clear pattern in the ACF, as every 12 lags (months) there is a significant ACF, suggesting we will have to include a seasonality parameter in the model.

In addition, we will need to work on the PACF, as there are multiple times where the PACF is significant.

We know that we are going to have to include a p and q parameter, of at least one, for both of them. If we run a ARIMA (1,1,1), we get the following residuals:

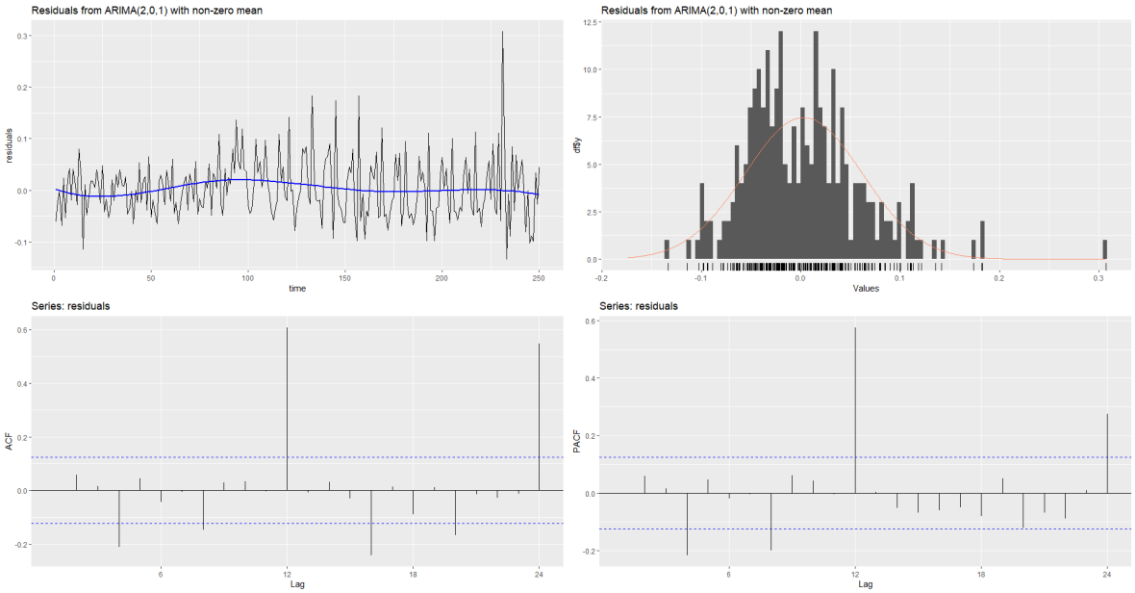



It is clear that, even though we have differentiated the ts, we still need to do something about the seasonality, as the ACF shows a trend, every 12 months. In addition, the PACF has not improved by using an AR(1) model.

The following model we fit will include the seasonality component, as we aim to eliminate it from our ACF.

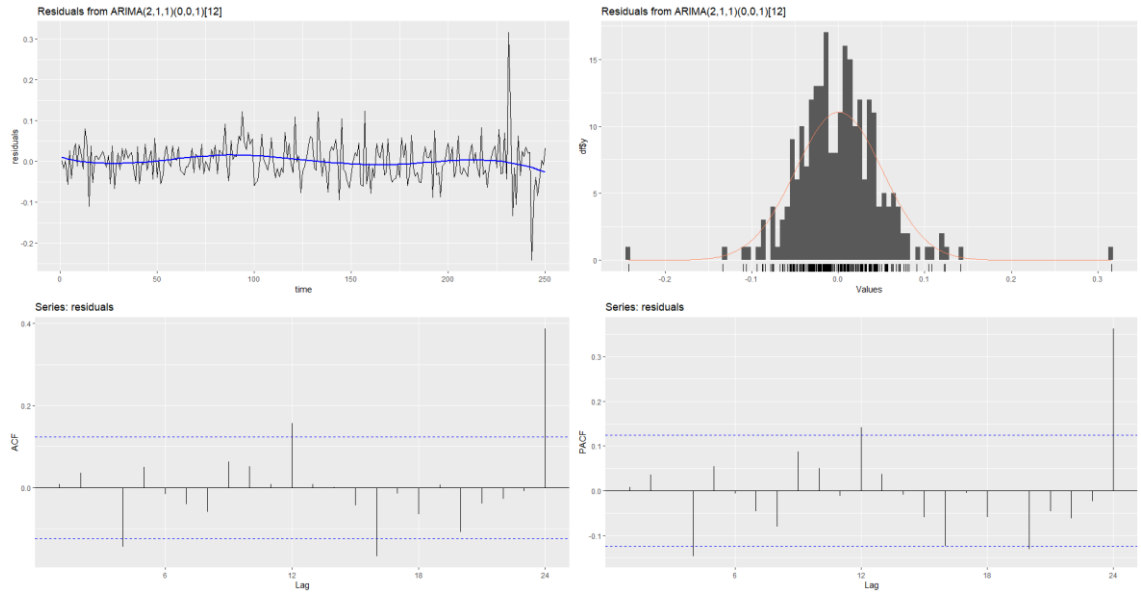
Before running a SARIMA model, we can check whether an ARIMA (2,1,1) would improve our model, as we still have some significant partial autocorrelations. Running said model, we obtain the following residual analysis:

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning




Although it seems that our ACF and PACF is worse, as we have much stronger significant autocorrelations, we observe that they occur every 12 lags, although we have eliminated any other significant autocorrelations. Seeing as we are going to include seasonality into the ARIMA model, this is not a problem for us. We expect, once we run the SARIMA model, that this seasonal trend does not appear.

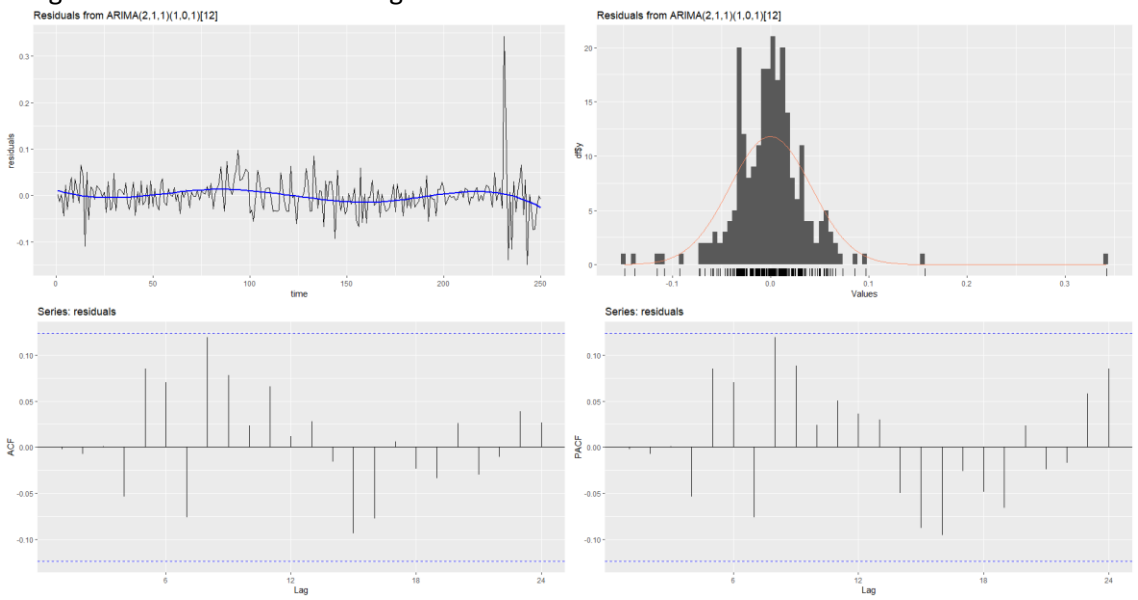
Once we have our ARIMA model (ARIMA (2,1,1)), we introduce a seasonality component. Given that we saw that the ACF still is seasonal, we expect to include a q term equal to at least one. But if we run the model, we still have some significant correlations.




We will try to deal first with the PACF, including in our latest model a p value equal to one.

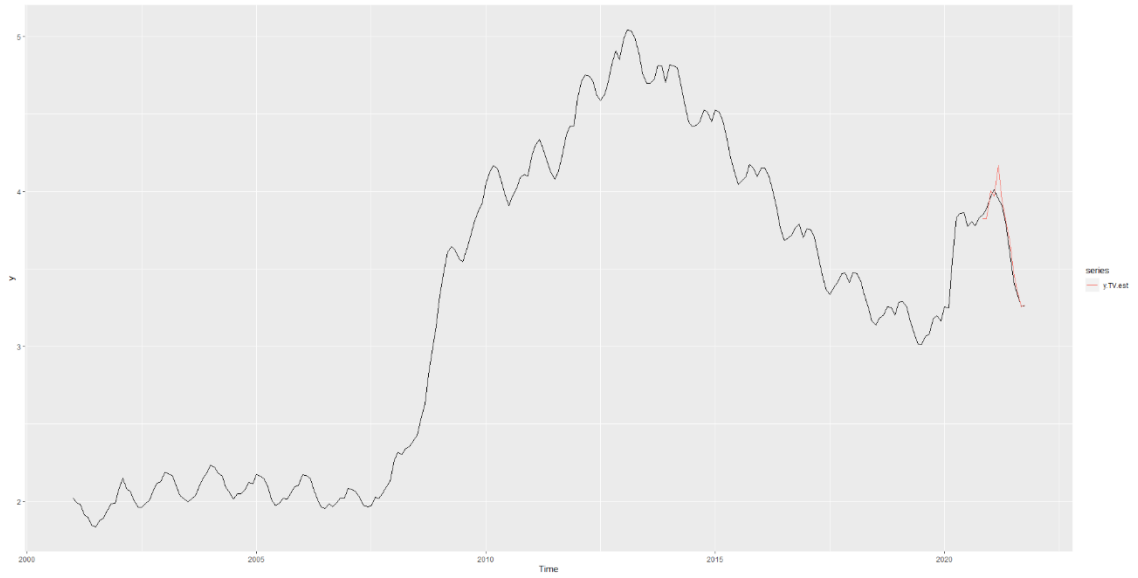
We get the following results:

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning




We can observe that we no longer have any autocorrelation present, suggesting that both the ACF and PACF are white noise (something we cannot model). In addition, the residuals are normal (except from some outliers which we will model with a dynamic model). We can also observe that the residuals (our error) increase substantially close to time = 250. This error in our modelling is caused by the Covid-19 pandemic, as suddenly a lot of people were laid-off, increasing the total unemployment rate.

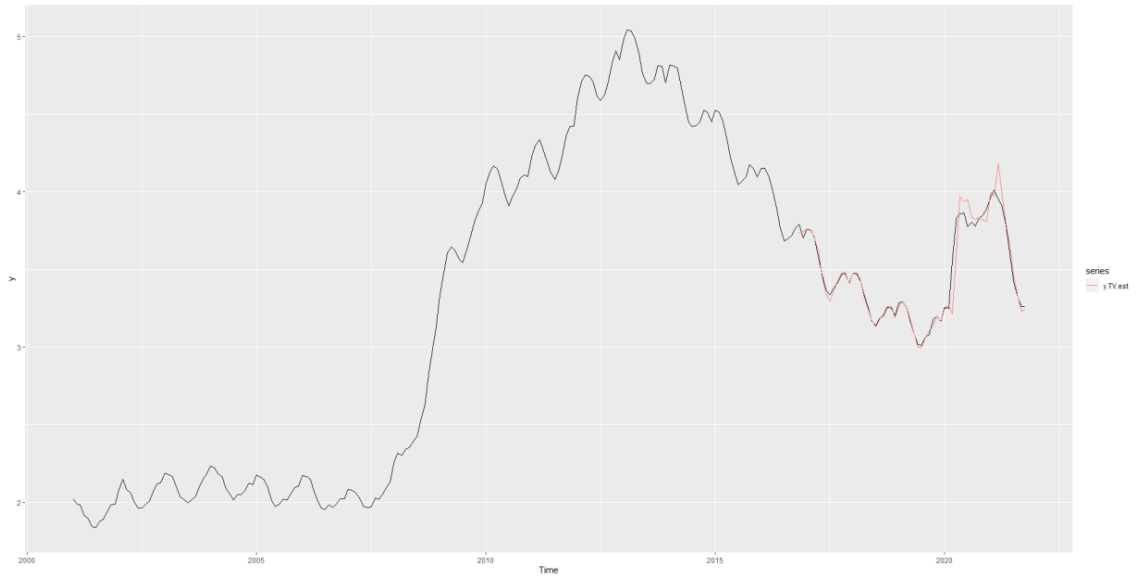
We can now proceed to check how well our test set will fare against the actual unemployment rate:




We observe that, at the start of our prediction, we overestimate the total unemployment number, but in the recent months we are much more accurate than before. This is likely due to the fact that we have chosen our test set to be the last year only, therefore we have not really considered covid, as it was already in full force.

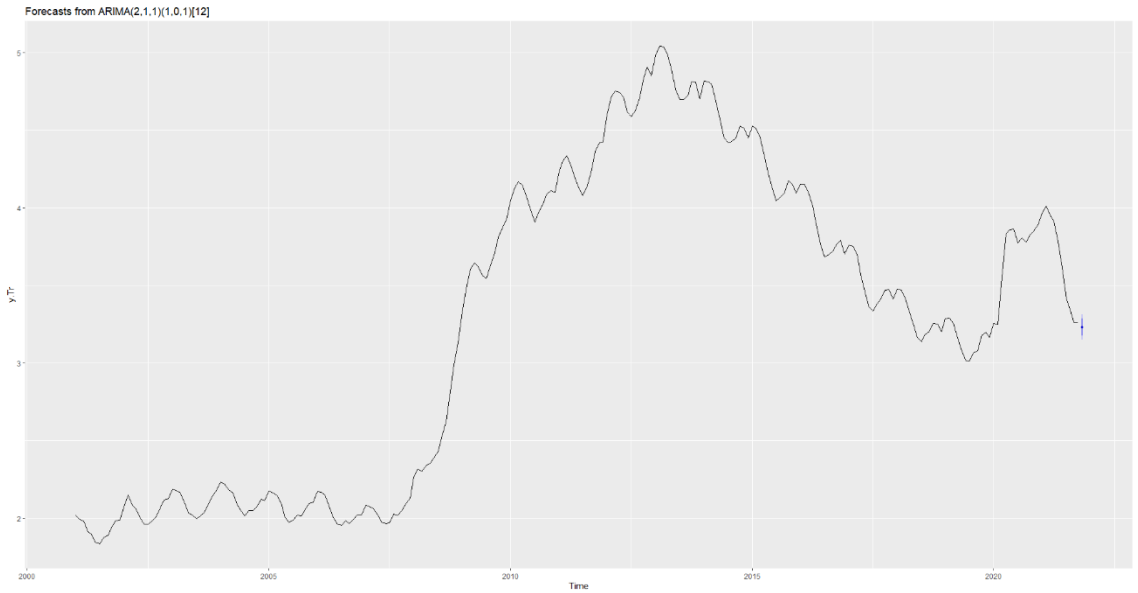
Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

If we choose a different time horizon for our test set (for example 5 years), we expect to have a larger error:




We are able to model very well the years prior to covid, but as soon as covid hits, are errors increase exponentially.

Future forecast:




| | Point Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 |
|---|---|---|---|---|---|
| Nov 2021 | 3.229939 | 3.17622 | 3.283659 | 3.147783 | 3.312096 |

We can see that our SARIMA model forecasts total unemployment to be around 3,229,939 million in Spain. We can compare this number to the actual number, 3,182,687. Although the point forecast is relatively far away from the actual number, we can observe that the actual

number was in our 80% confidence interval (very close to our lower bound). Therefore, we can conclude that our SARIMA model can accurately predict the unemployment number in Spain. We can also predict what the unemployment number will look like in December.

| | Point Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 |
|---|---|---|---|---|---|
| Nov 2021 | 3.229939 | 3.176220 | 3.283659 | 3.147783 | 3.312096 |
| Dec 2021 | 3.187618 | 3.085575 | 3.289661 | 3.031557 | 3.343679 |

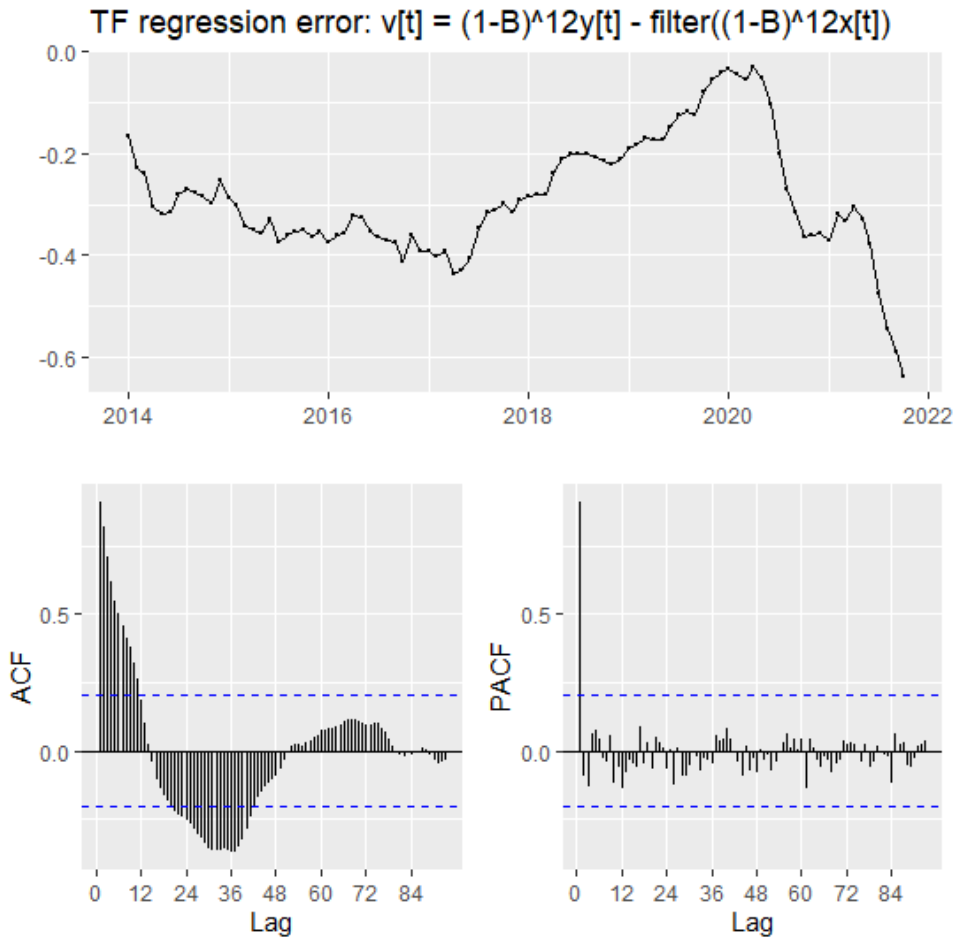
Our model predicts again that the total unemployment will lower again, to pre-pandemic levels. Our 80% CI has a lower bound, compared to November, while it has the same upper-bound.

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

## Dynamic Regression model

In order to implement a Dynamic Regression, we decided to establish a new variable called COVID, that represents with 0 and 1 the time where the COVID had more impact on the unemployment rate, this variable is needed because as this disease was a non-expected event, causing a break-point on the time series, which affects the results of our forecasting model, because they are strong trend changers, at the same way we decided to filter the years, using a windows in order to take just values from the year 2013 forwards.

In the first we have the results of our first training of the model, as we can see clearly that the regression error is not stationary, taking in count that variance is not stables or the mean value, and obviously the trend that is showing us in the ACF plot, for this output we took values or order 1 and the standard values of seasonal, and values of r and s, of 0 for r, and 9 for s, just to take in count for this first iteration of the model we took all the values of the dataset.

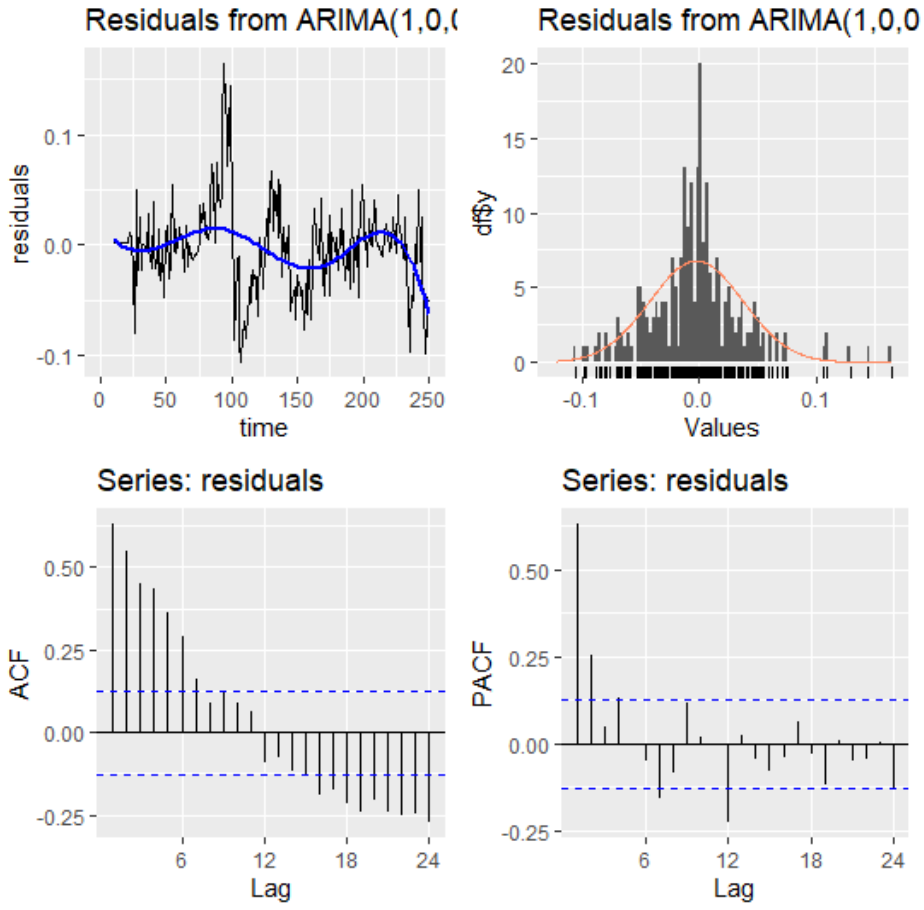


Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

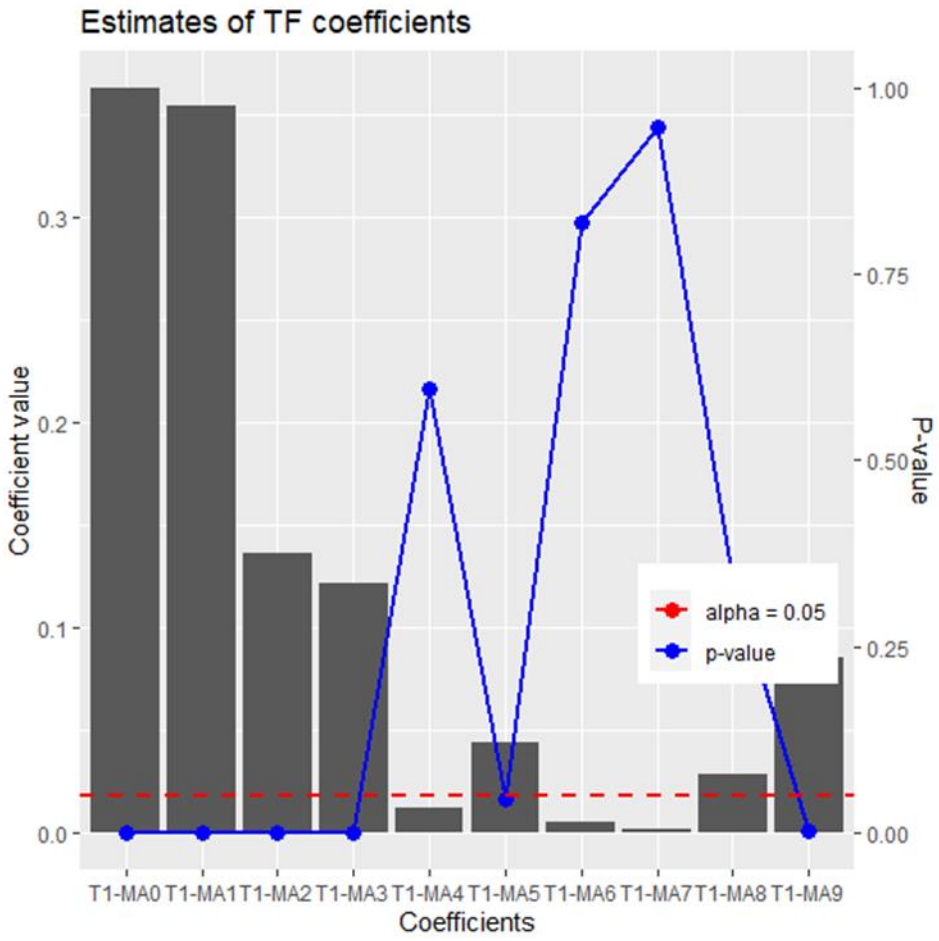
For this first iteration we can see in the following next two plots, we can see that the performance of this prediction is hardly the best, considering we had a p-values of 2.2e-16, multiplied values of correlation where it supposed to be white noise, and just a few of representative values taking in count that it should be all of them, but the secund plot give us the next values for our r and s.

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,1,0)[12]
Q* = 474.55, df = 13, p-value < 2.2e-16

Model df: 11.   Total lags used: 24
```

```
z test of coefficients:

         Estimate Std. Error  z value  Pr(>|z|)
ar1    0.99534565 0.00486489 204.5977 < 2.2e-16 ***
T1-MA0 0.34878266 0.02892324  12.0589 < 2.2e-16 ***
T1-MA1 0.34811914 0.02892325  12.0360 < 2.2e-16 ***
T1-MA2 0.13326629 0.02892327   4.6076 4.074e-06 ***
T1-MA3 0.12042113 0.02892328   4.1635 3.135e-05 ***
T1-MA4 0.01116001 0.02892330   0.3858   0.69961
T1-MA5 0.04384586 0.02892332   1.5159   0.12954
T1-MA6 0.00473747 0.02892334   0.1638   0.86989
T1-MA7 0.00088448 0.02892336   0.0306   0.97560
T1-MA8 0.00165787 0.04093972   0.0405   0.96770
T1-MA9 0.06832827 0.04093884   1.6690   0.09511 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

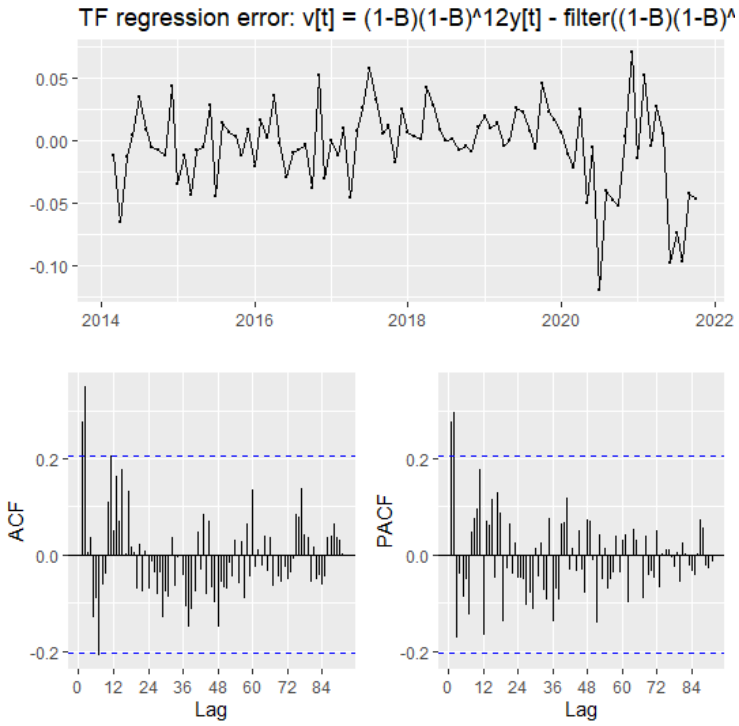


Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning
Analyzing this plot, we can notice that the best values for our r and s will be 1 and 1 respectively.




Now for this secund iteration we took only the values from 2013 and forwards, we took the vales obtained from the past iteration, where it said that the values of r and s should be 1, so in this case we add one differentiation and order 1, and a seasonal of (1,1,0) to analysis the performance of the model, using these values we can appreciate that the model had improved considerably, noticing that we have a better distribution of the regression error in variation and mean but still can we improved and mor if consider the p values that we had that is just of 0.0015, and if we analysis the coefficients we can noticed that all of them are representative as it has to be, but considering the values of ACF and will train to more iteration and we will keep the best of those two.

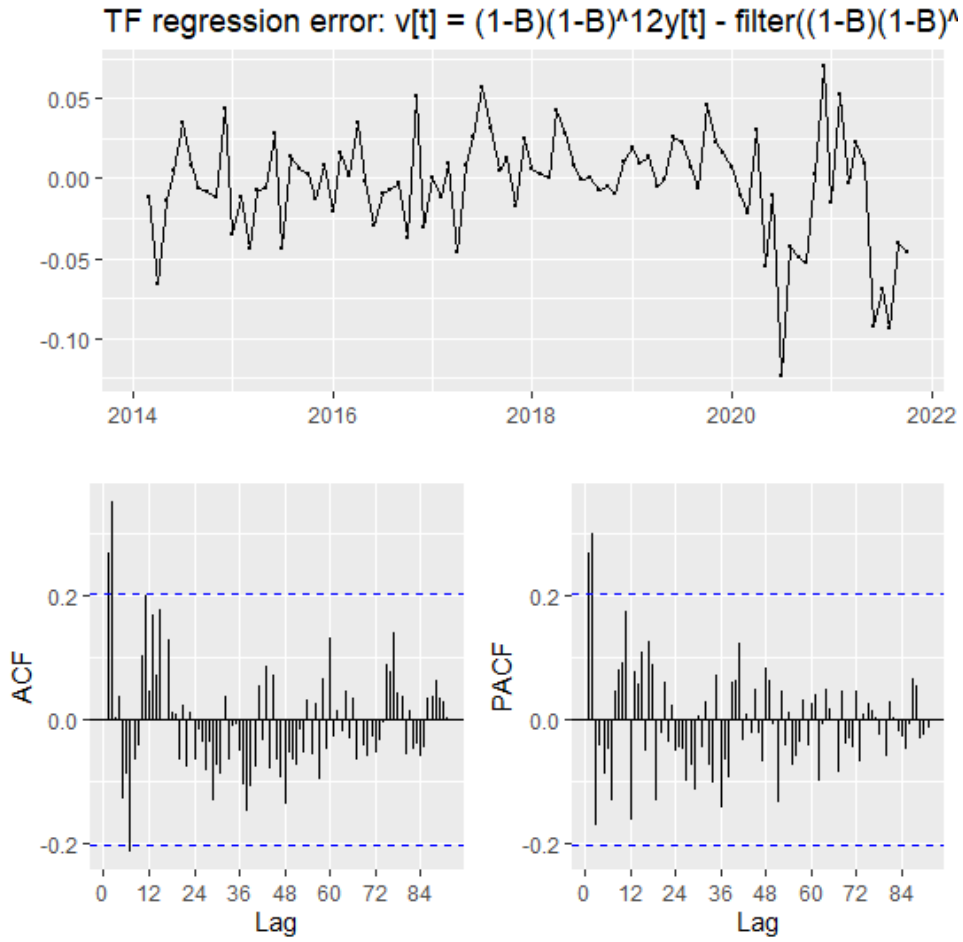
Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning




For the next model we are considering the values in the position 2 as the order is almost inside of the area of white noise, for ACF and PACF we establish the same criteria.

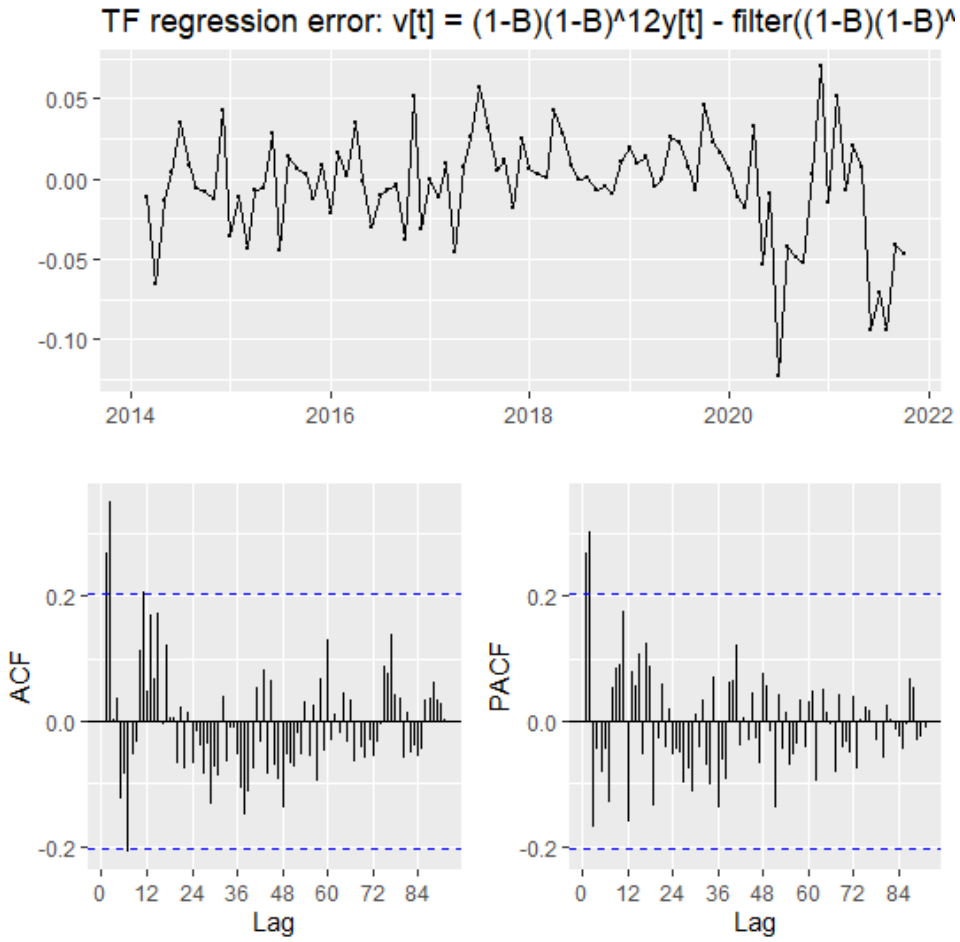
Now in the case taking a value of order 2 and 1 differentiation, the same stationarity, we got better result than the previous model with a p-value of 0.2749, a great number of relevant variables, and the result for the residuals the considered is acceptable we are just torching the limit of the white noise, and we think is better to have the model this way, because is easier to explain in the case that we had to.

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning



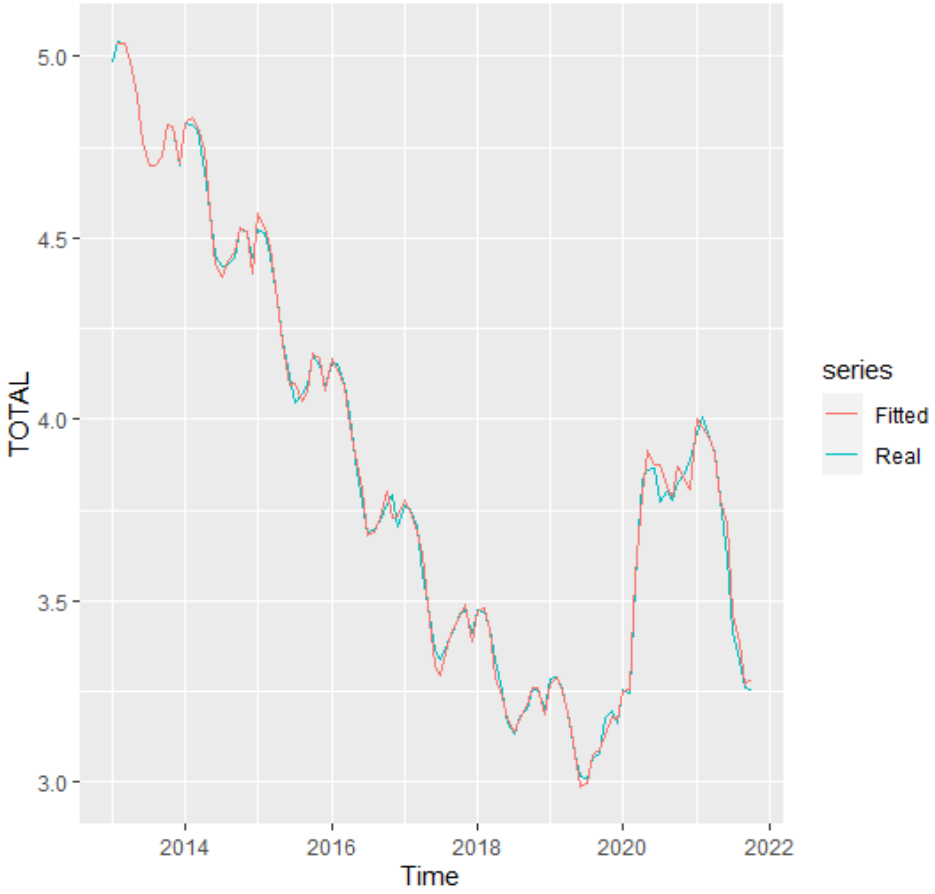

In this last iteration we took the values that que had before in the PACF that's why we took order c (0,1,2), and the same stationary, we didn't notice a huge difference between these two models, the p-value change to 0.2661, but for this particular case we would take the previous model, because it has a better distribution of white noise

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning




And now that we have our final model the one of a p-value of 0.2749, we can see how the model is fitting the forecast, in our point of view the model have a good performance, just a little variance in the range of covid and as is a break point is something expected.

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning



Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

## Model Comparison

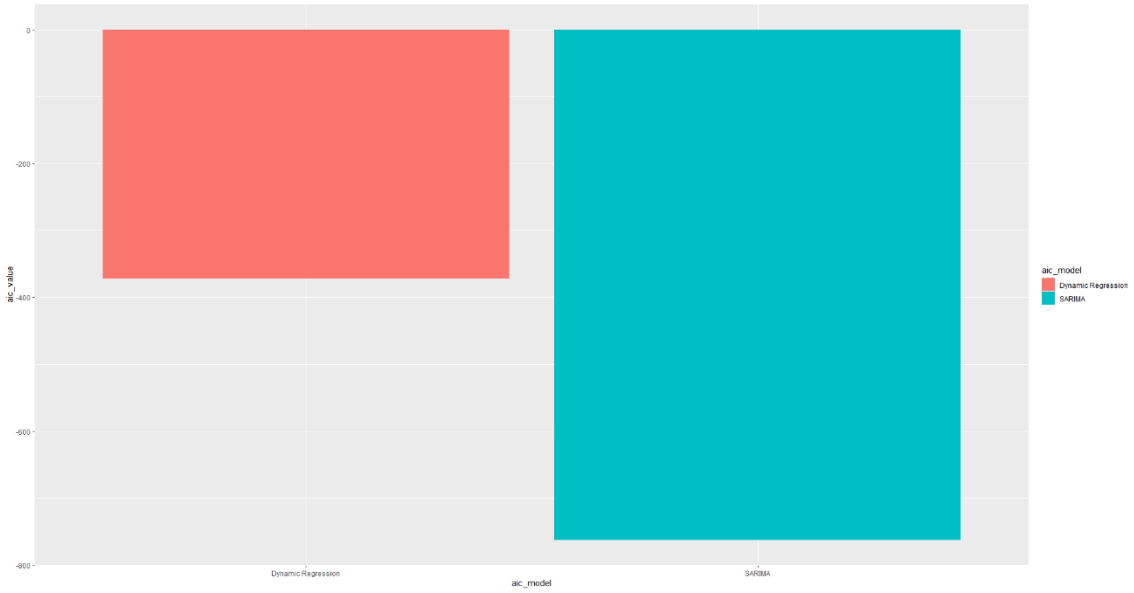
Now that we have two different approaches to forecast the absolute unemployment in November for Spain, we can proceed to compare both models.

Firstly, we will look at the AIC value for both models. The AIC is a mathematical method to evaluate how well a model fits the data. Given that both of our data was in the same units (i.e., we scaled both of the into units), we can directly compare the AIC values. The lower the AIC value is, the better the model fits the original data

The following table is our AIC values for both models:

| | aic_value |
|---|---|
| Dynamic Regression | -372.5042 |
| SARIMA | -762.8367 |

We can see that, the SARIMA model has a lower AIC, meaning that the model fits the original data better than the dynamic regression.

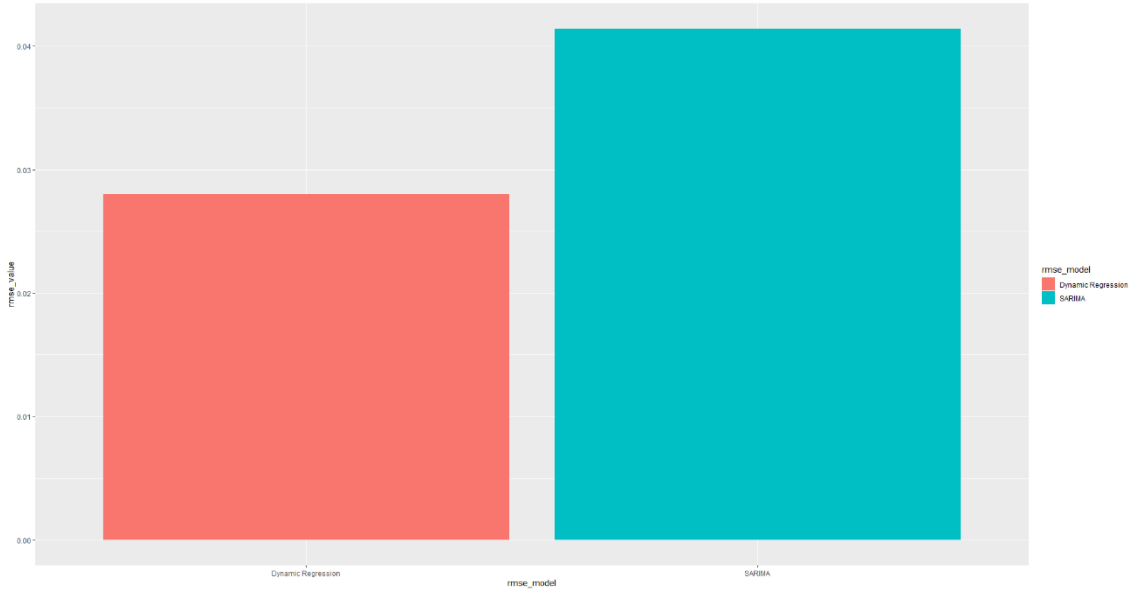



But we cannot conclude that the SARIMA model is better than the dynamic regression. We need to look at more than one metric to conclude what model is best. We now look at the root mean squared error, which measures the deviation of the predicted values compared to the actual values. In this case, we are looking for the smallest number.

| | rmse_value |
|---|---|
| Dynamic Regression | 0.02803065 |
| SARIMA | 0.04141120 |

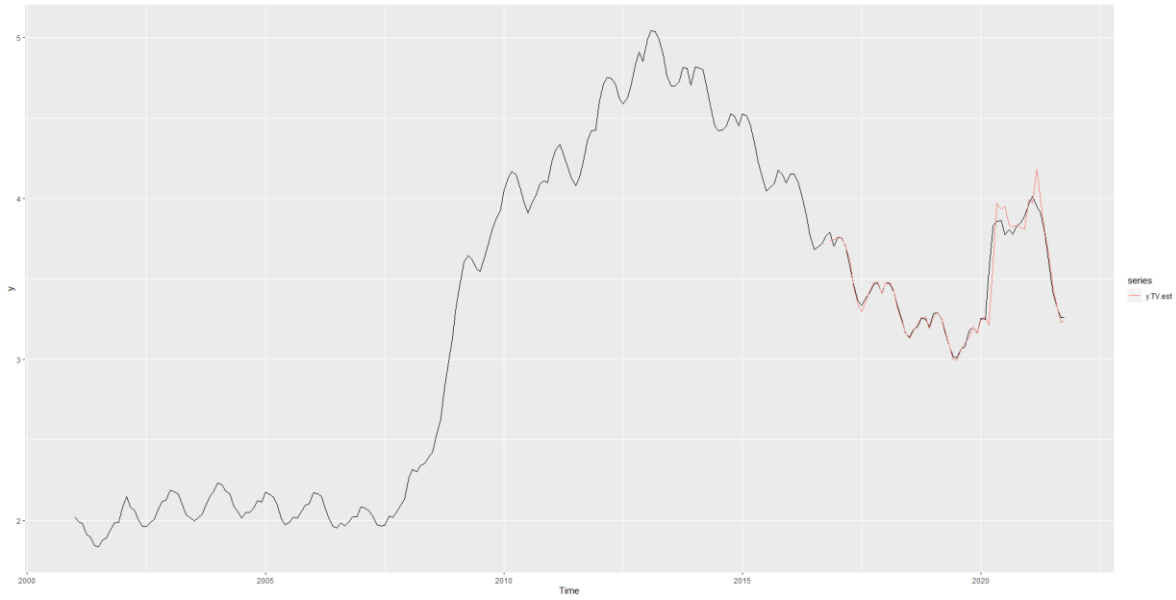
We can observe, that in this case, the dynamic regression model substantially out-preforms the SARIMA model.

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning




We can also visually compare how well our models fit the data, and the one that most closely resembles the given data could be considered the best model.

SARIMA:



Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning
Dynamic regression with Covid intervention variable:


From the above graphs, we can see that the dynamic regression model is a better fit compared to the SARIMA model. Both models closely resemble the data up until 2020 (when covid hits). But when trying to model the impact covid has, the SARIMA model is always lagged, as we did not introduce a variable to model this unprecedented situation.

Larry Linares
Nicolas Garcia
Assignment 2 – Machine Learning

## Final prediction

3,229,939