STA 137 Final Project

Qinyi Qiu
6/5/2022

Forecasting the Export of Central African Republic Export

I. Introduction & Background

The Central African Republic is a country located in the center of Africa. The development of the economy of the Central African Republic relies heavily on exportation, especially products like timber, diamonds, cotton, and coffee. Hence, we are interested in the exportation of the Central African Republic. In this study, we try to predict the exportation in the Central African Republic after 2017 by analyzing the historical exportation data from 1960 to 2017 using a time series model. In order to better forecast the exportation, we decide to build the ARIMA model, a statistical model for analyzing and predicting time series data, for finding out a better fit model and forecast the exportation after 2017.

II. Summary of Data

In this study, we use the dataset Central African Republic, which includes 9 variables (Country, Code, Year, GDP, Growth, CPI, Import, Export, Population) and the sample size is 58. In this study, as we are interested in only exportation, we only focus on the Export variable (percentage of exports of goods and services in GDP), with 58 sample sizes from the year 1960 to 2017.

Country <fctr>	Code <fctr>	Year <dbl>	GDP <dbl>	Growth <dbl>	CPI <dbl>	Imports <dbl>	Exports <dbl>	Population <dbl>
Central African Republic	CAF	1960	112155599	*NA*	*NA*	34.18181	23.27272	1503508
Central African Republic	CAF	1961	123134584	4.9535538	*NA*	35.76159	26.49007	1529227
Central African Republic	CAF	1962	124482749	-3.7138002	*NA*	37.70491	24.59017	1556661
Central African Republic	CAF	1963	129379098	-0.7070108	*NA*	38.48581	25.23659	1585763
Central African Republic	CAF	1964	142025069	2.0803246	*NA*	40.80459	28.44827	1616516
Central African Republic	CAF	1965	150574816	0.9475787	*NA*	37.66938	27.10027	1648833

The range of our data is between 10.68442% in the year 2009 and 34.31230% in the year 1968. That means from the year 1960 to 2017, the Central African Republic had the highest 34.31230% exportation of GDP in the year 1960 and the lowest 10.68442% exportation of the GDP in the year 2009.

According to the plot Export V.S. Year, we can see there exist an obviously trends in the changing of export over the time increase. The plot shows that in the early years, from about the 1960s to the 1970s, the rate of exportation to GDP approximately fluctuated between 25% and 35%. After the year the 1970s, we can see that the polyline in the plot shows an obviously downward sloping, dropping from about 30% (1970s) to 10% (2010s), which means that after the 1970s, the ratio of export to GDP kept decreasing.

III. Diagnostics

In time series analysis, we predict the future data by assuming that each data point is independent and stationary, having the same statistical properties and constant behavior for each data point over time. Only with a dataset that has a constant overall behavior, we can use a time series model to forecast the future model.

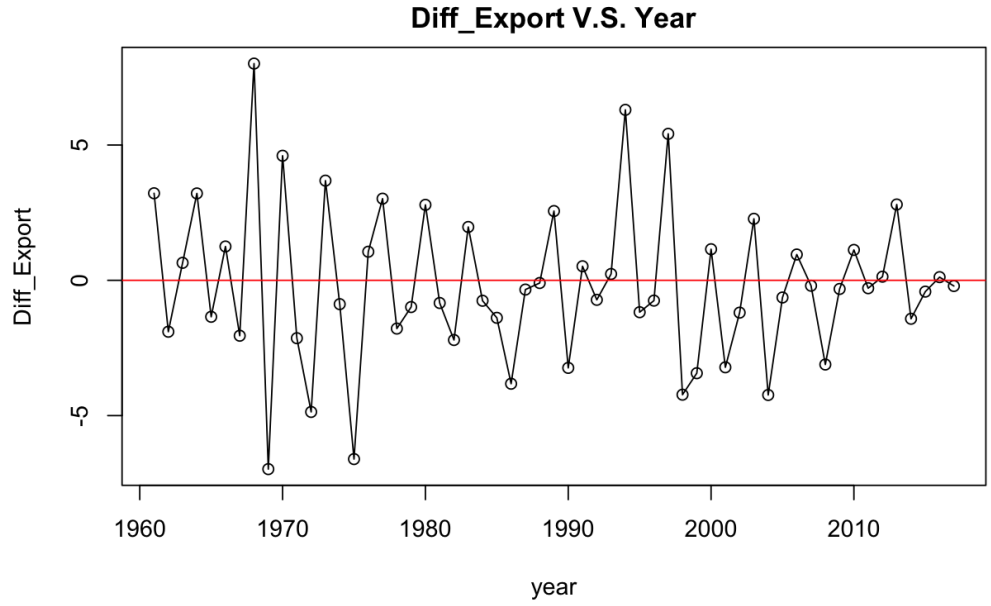
Stationary

As we mention above, stationary is an important assumption in the time series analysis. Stationary means that the statistical properties of the data do not change over time. For example, a dataset with an obviously increasing or decreasing trend over time is not stationary since the behavior of the data keeps changing over time.

According to the Export V.S. Year plot, our dataset is not stationary since it has an obviously decreasing trend. Hence, we cannot directly use time series analysis for our data forecasting, as our dataset violates the assumption of time series analysis. In order to apply time series analysis for our forecasting, we need to make a transformation to stabilize our data, making it becomes a stationary dataset.

Differencing

Differencing is a transportation method that tries to make nonstationary data to stationary by computing differences between two consecutive data points ($Y_t - Y_{t-1}$), helping to reduce the trend of the data. We use differencing transportation in our dataset.



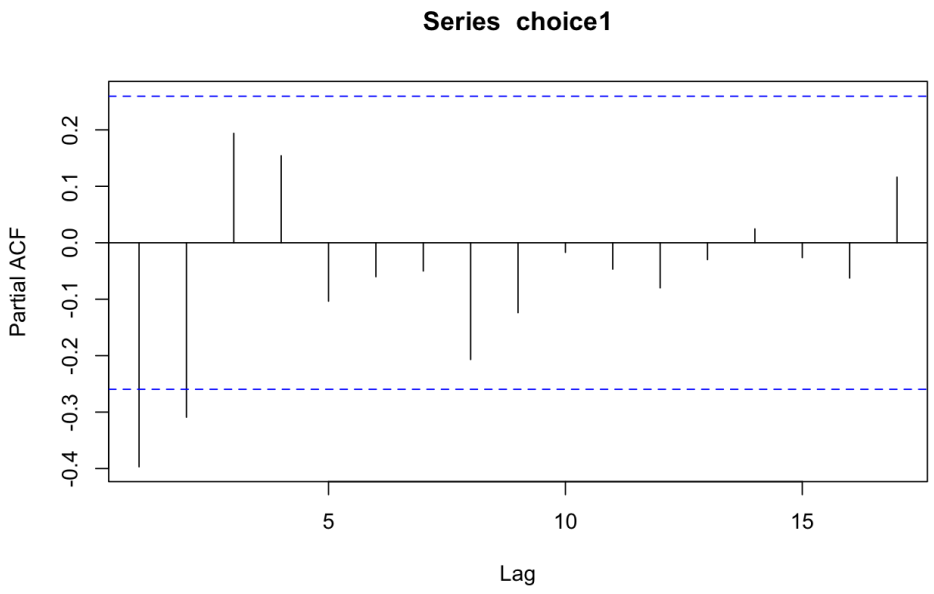
The plot Diff_Export V.S. Year shows that, after differencing transportation, our data do not show any apparent trend or pattern. We can say our dataset now is stationary and does not need further transportation. Hence, we use data after differencing transportation in our following time series analysis.

IV. Analysis

With our stationary data, we look for candidate forecasting models for our data by fitting our dataset into different time series models such as the autoregressive model (AR), and the Moving Average model (MA)

AR model:

Using observations (ex. Y_{t-1}) from the previous time step as the input in a regression equation to predict the value in the next time step. To figure out how many lagged time steps we need in our prediction, we need to look at the PACF plot for significant lagged time steps. PACF is a plot finding correlation coefficient between residual and lagged value. If we find that there exists a high correlation between the residual and a lag time, exceeding the confidence interval of correlation, we should include the lagged time in our regression equation.

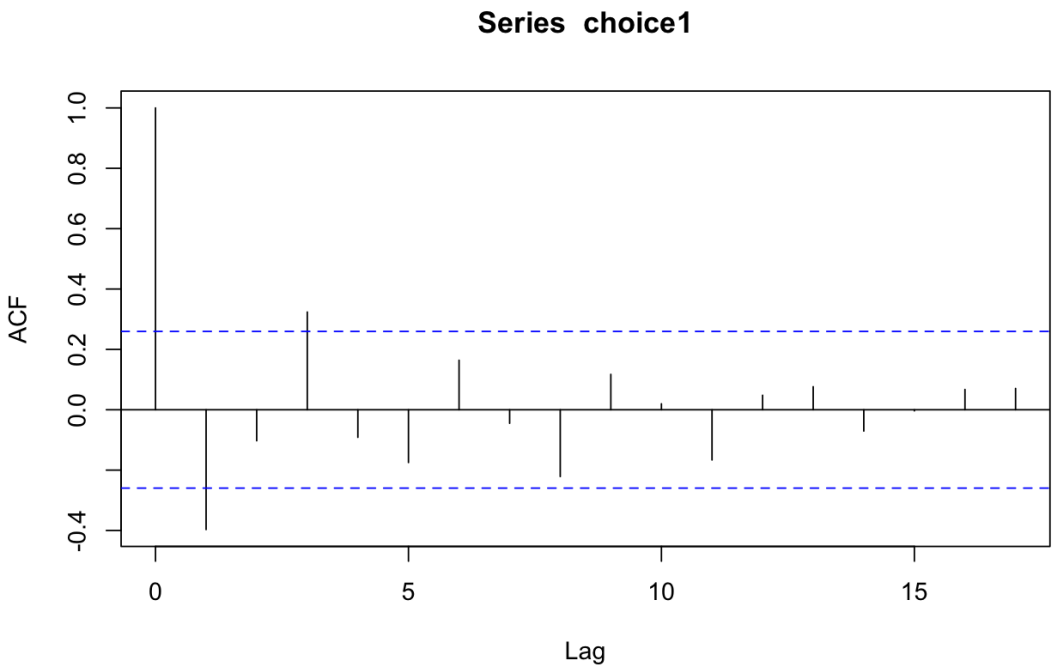


According to the PACF plot of our data, the x-axis presents the lag time and the y-axis presents the correlations between lagged value and residual. We can see the first two lagged values have a high correlation with the residual, which is beyond the blue dotted line. That means the first two lagged observations are significant in our regression equation. In conclusion, we get our first candidate model AR(2) for our forecasting.

MA model:

Using past forecast error as input in regression equation to predict present value. To figure out how many lagged values we should obtain, we look at the ACF plot, a plot that shows the correlation between the present value and lagged value (ex. Correlation

coefficient between Y_t and Y_{t-1}). We should contain lagged values that ACF beyond the confidence interval.



According to the ACF plot, we learn lag 3 is beyond the upper confidence interval of the correlation coefficient. Hence, we obtain 3 lagged values in our regression. We have our candidate model MA(3).

In order to pick a better model from our candidate model (AR(2) and MA(3)), we look at their p-values and compare their AIC and BIC. The model with a smaller AIC and BIC is a better predictor.

AR(2)

```
$ttable
      Estimate     SE t.value p.value
ar1    -0.5230 0.1262 -4.1460  0.0001
ar2    -0.3065 0.1248 -2.4563  0.0173
xmean  -0.2120 0.1841 -1.1514  0.2546

$AIC
[1] 4.829009

$AICc
[1] 4.836953

$BIC
[1] 4.972381
```

MA(3)

```
$ttable
      Estimate     SE t.value p.value
ma1    -0.4537 0.1319 -3.4387  0.0011
ma2     0.0922 0.1532  0.6018  0.5499
ma3     0.2677 0.1354  1.9762  0.0533
xmean  -0.1999 0.2946 -0.6787  0.5003

$AIC
[1] 4.838539

$AICc
[1] 4.852034

$BIC
[1] 5.017754
```

The p-value of AR(2) shows that the first lagged time step and the second lagged times are significant in the regression equation, as all p-values smaller than 0.1, and the AIC

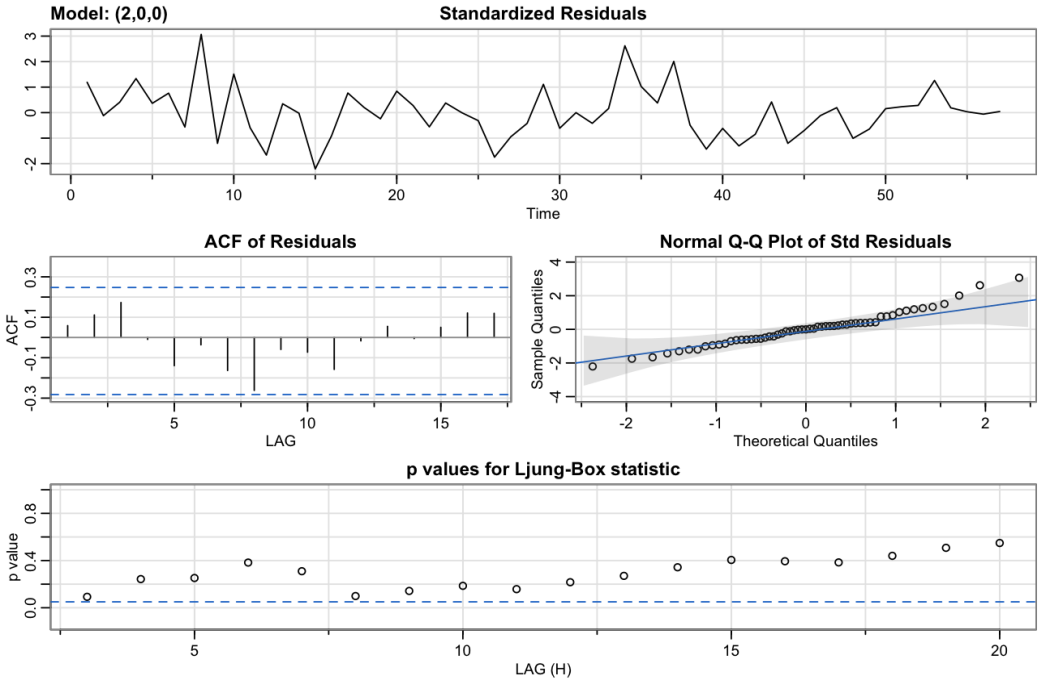
equals 4.829, BIC equals 4.972. The p-value of MA(3) shows that the second lagged time is not significant, as the p-value in ma2 is 0.5499 which is larger than 0.1. Hence, we decide not to include nonsignificant lag time in our prediction and change the candidate model to MA(1), which only includes one significant lag time step.

MA(1)

```
$ttable
      Estimate     SE t.value p.value
ma1    -0.4431 0.1077 -4.1149  0.0001
xmean  -0.2141 0.1945 -1.1006  0.2759

$AIC
[1] 4.856507

$AICc
[1] 4.860406

$BIC
[1] 4.964036
```

The AIC for MA(1) equals 4.8565, and BIC equals 4.964. Comparing the AIC and BIC of AR(2) with MA(1), we find that both models have approximately the same AIC and BIC. Hence, it's difficult to choose a better model from them by looking at AIC and BIC. Then we check the residuals for our candidate models by plotting the ACF of residuals and doing a portmanteau test. If our models look like white noise, then our model should be a good predictor for our data.

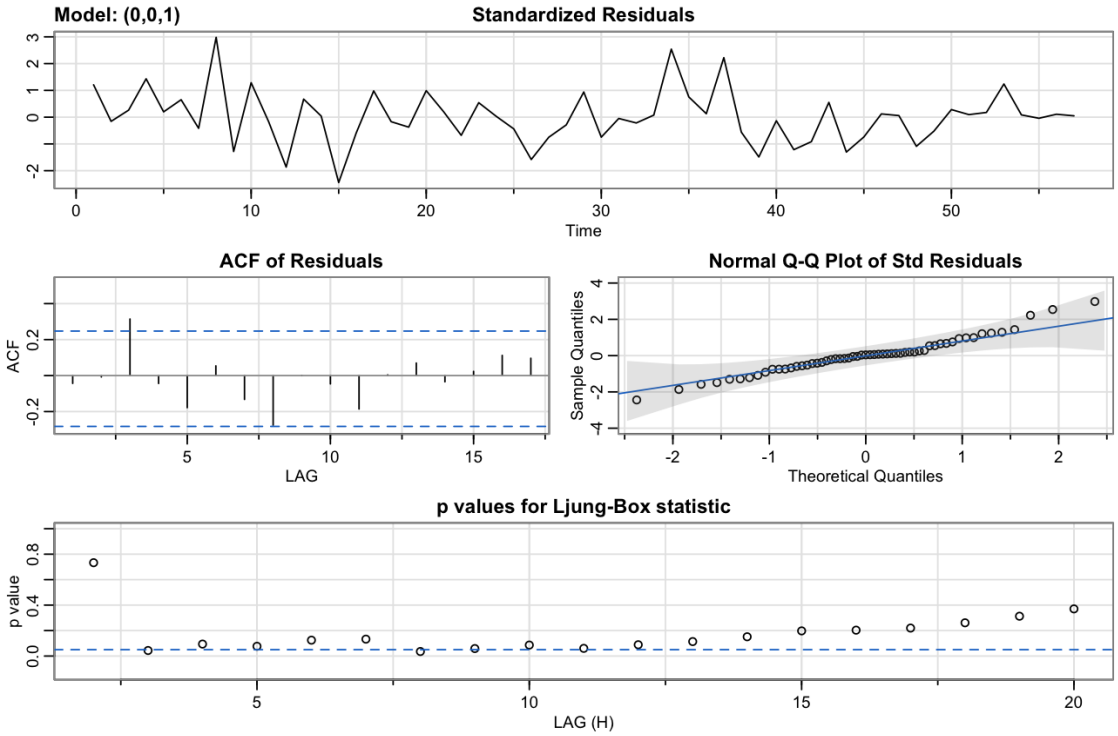
AR(2)



From the plot of the standardized residuals, we can see there are no obvious trends and patterns, which means, residuals are independent, random, and stationary. From the ACF

of the residuals plot, we can see that there is no high correlation coefficient so it is white noise. We apply the portmanteau test, with the null hypothesis: the residual is white noise. And use the p-value to test the null hypothesis. Since we know from the p-value for the Ljung-Box statistic plot that all our p-values are greater than 0.1, we fail to reject the null hypothesis and conclude that the residual is white noise.

MA(1)

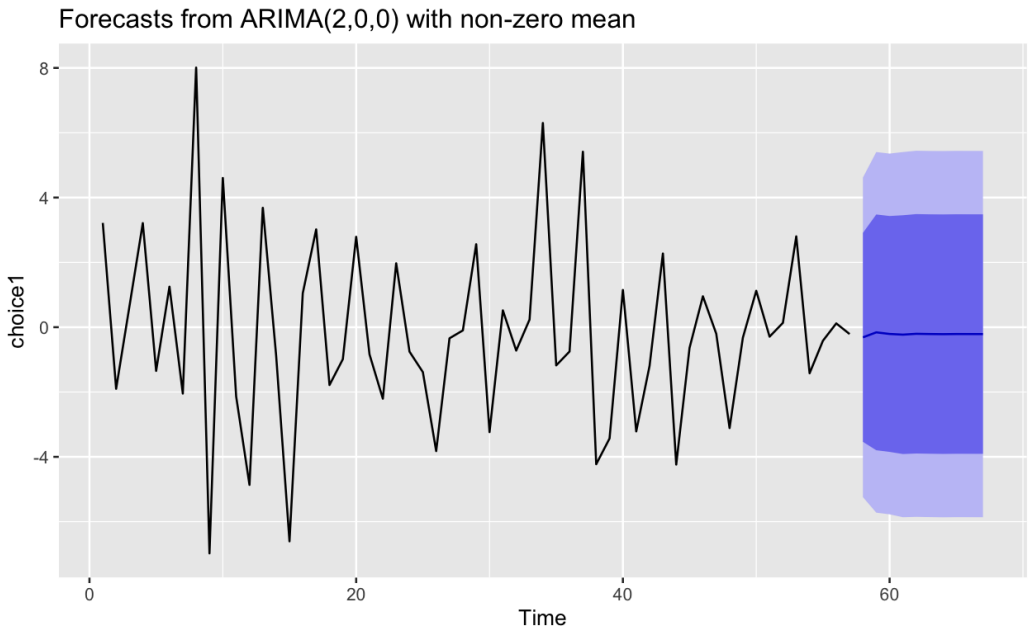


From the plot of MA(1), we learn that its residuals also have no trend or a specific pattern. However, in the ACF plot, there exists a high correlation coefficient between residuals. Furthermore, applying the portmanteau test, a few p-values are smaller than 0.1. That means we may reject the null hypothesis at some points. We may not say residuals for this model are white noise for sure.

V. Interpretation

By comparing AIC, BIC, and the residual plot for both candidate models, we decide that AR(2) is a better predictor for our dataset. Both AIC and BIC are extremely close in AR(2) and MA(1). Under the situation, since we are more sure about the residuals of AR(2) are white noise compared with MA(1), we prefer AR(2), though MA(1) is an easier model.

Then we apply AR(2) for our forecast. According to the coefficients we get above, AR(2) can be written as: $Y_t = -0.523*X_{t-1}-0.3065*X_{t-2}+W_t$



	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
58	-0.3133799	-3.536014	2.909254	-5.241973	4.615213
59	-0.1592450	-3.796054	3.477564	-5.721264	5.402774
60	-0.2084660	-3.846826	3.429894	-5.772858	5.355926
61	-0.2299685	-3.913053	3.453116	-5.862760	5.402823
62	-0.2036345	-3.896364	3.489095	-5.851177	5.443908
63	-0.2108169	-3.903721	3.482087	-5.858626	5.436993
64	-0.2151324	-3.909405	3.479140	-5.865034	5.434770
65	-0.2106736	-3.905180	3.483833	-5.860933	5.439586
66	-0.2116829	-3.906200	3.482835	-5.861960	5.438594
67	-0.2125217	-3.907081	3.482037	-5.862862	5.437818

The plots show the prediction of differences between two consecutive observations after 2017 (ex. Difference in export ratio to GDP in January 2017 and February 2017). For example, we are 80% confident that the value of the next difference after 2017 lies within the intervals between -3.536 and 2.909. And we are 95% confident that it lies within the intervals -5.242 and 4.615. The specific difference value we predict is -0.3133. Since the difference is defined as the present value (Yt) minus the previous value (Yt-1). That means we predict the present export ratio (Yt) is larger than the previous export ratio (Yt-1). Notice that all our predicting differences are negative, we conclude that the future value of export ratio to GDP is downward slopping.

VI. Conclusion

In conclusion, using the ARIMA model, we find that AR(2) model is an appropriate prediction model for our data and we predict that all differences value between two consecutive observations in our data after 2017 is negative. We predict that the exportation ratio to GDP of the Central African Republic will decrease. However, our prediction may not very precise since we are only based on AR(2), more research can be done.

Appendix

```
load("/Users/erika/Downloads/finalproject.Rdata")
data<-finalPro_data
library(astsa)
library(forecast)
head(data)
plot(x=data$Year,y=data$Exports,xlab = "year",ylab = "Export",main = "Export V.S. Year",type = "o")
range(data$Exports)
choice1<-diff(data$Exports)
length(choice1)
plot(x=(1961:2017),y=choice1,xlab = "year",ylab = "Diff_Export",main = "Diff_Export V.S. Year",type
= "o")
abline(h=0,col="red")
acf(choice1)
pacf(choice1)


#Candidate model: MA2,AR2
sarima(choice1,2,0,0)#(AR2)
sarima(choice1,0,0,3)#(MA3)
sarima(choice1,0,0,1)#MA1

l=forecast(arima(choice1,order=c(2,0,0)))
autoplot(l)
l
```