About the Project:

Gross Domestic Product (GDP) is one of the most widely used measures of an economy's output or production. It is defined as the total value of goods and services produced within a country's borders in a specific time period — monthly, quarterly or annually.

GDP allows policymakers, economists and business to analyze the impact of economic shocks such as a spike in oil price, as well as tax and spending plans, on the overall economy and on specific components of it. Therefore, the analysis of GDP fluctuations during the time is mandatory for governments in order to predict the future condition of their economy.

Description of data:

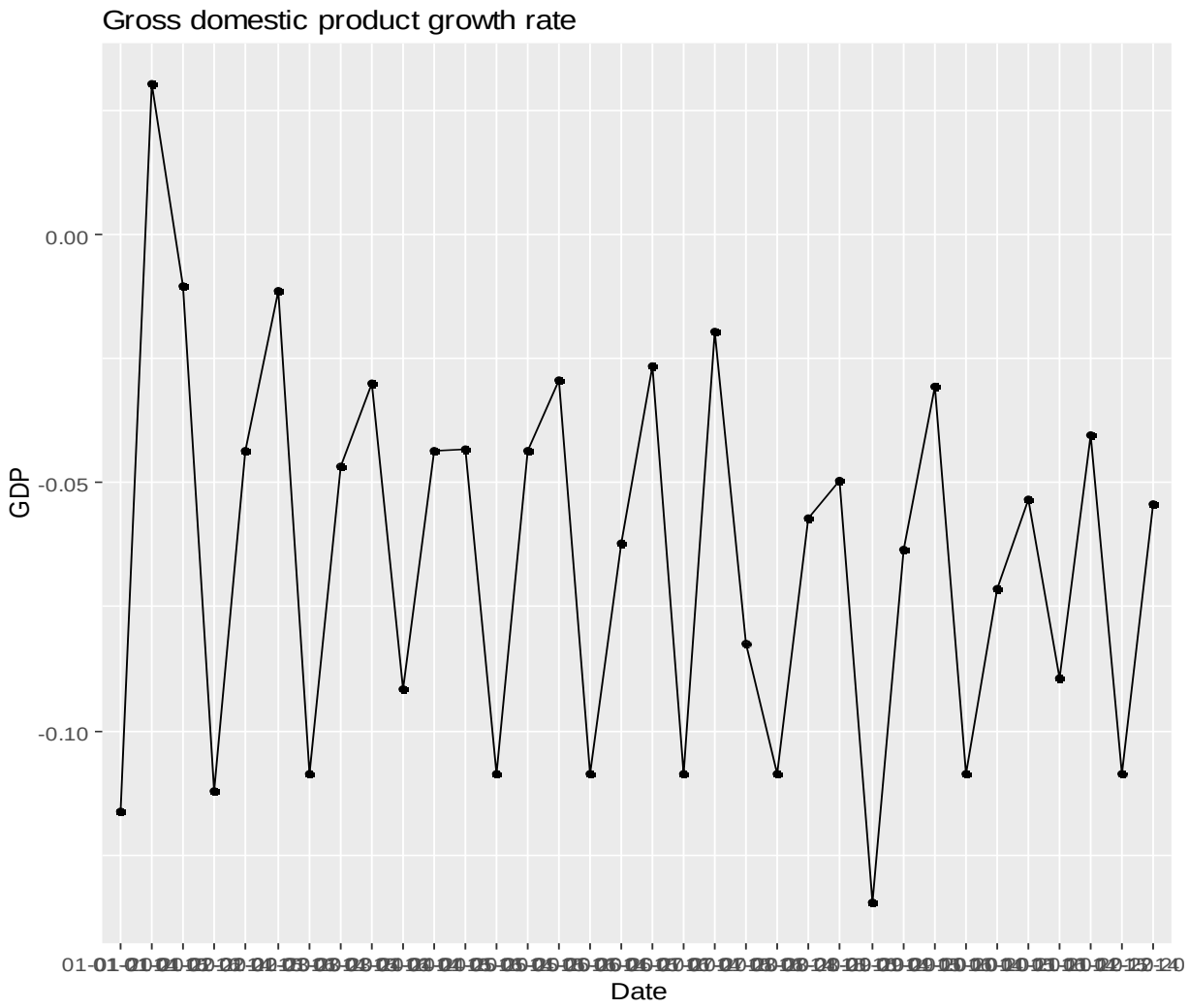
The data consists of 2 variables, Date as independent variable and Gross Domestic Product Growth Rate as dependent variable which measures how fast the economy is growing or declining. Data are collected monthly from Jan 2014 to Oct 2016 (at the beginning of every month) in the country of Iran.

Data Analysis Methods:

In this project non-seasonal ARIMA model, seasonal ARIMA and Linear Regression are used to figure out the probable trends in data and make predictions.

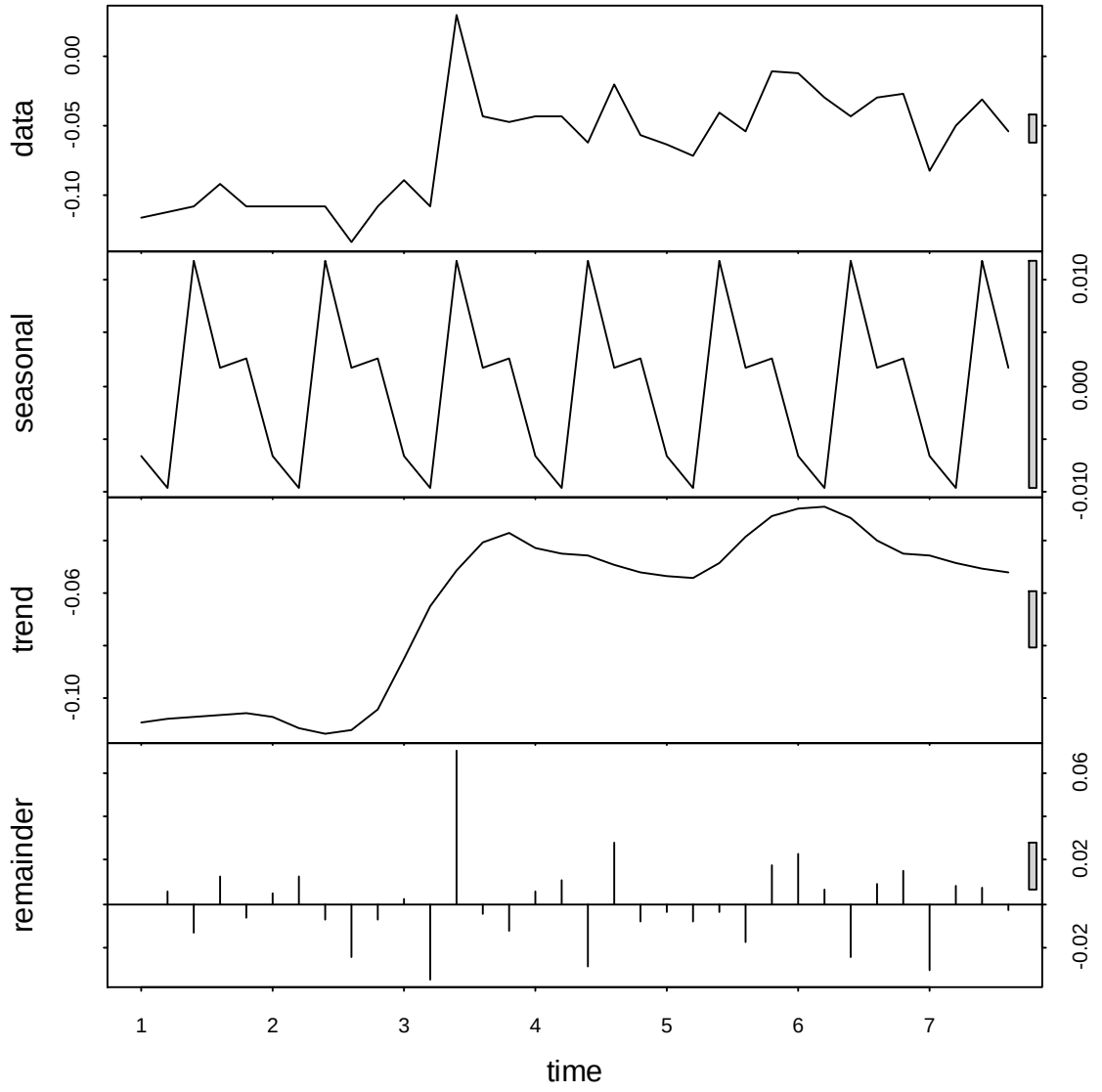
ARIMA stands for auto-regressive integrated moving average and is specified by these three order parameters: *(p, d, q)*. An auto regressive (AR(p)) component is referring to the use of past values in the regression equation for the series *Y*. The auto-regressive parameter *p* specifies the number of lags used in the model. The *d* represents the degree of differencing in the integrated (*I(d)*) component. Differencing a series involves simply subtracting its current and previous values *d* times. Often, differencing is used to stabilize the series when the stationarity assumption is not met. A moving average (MA(q)) component represents the error of the model as a combination of previous error terms.

As the first step we plot the series and some patterns are showed in the following diagram:



Analysis of seasonality, trend, and cycle (decomposing the data)

In the next chart it is represented how the series is decomposed and the seasonality is removed using subtracting the seasonal component from the original series.



Analysis of stationarity

A stationary process is a stochastic process whose parameters such as mean and variance do not change over time. Before fitting ARIMA model to the data we need to check the stationarity of the data.

Here, the null hypothesis refers to nonstationarity and alternative hypothesis is otherwise. Based on the output in R, p-value is big (more than 0.05 for 95% confidence interval) and we don't accept the alternative hypothesis and series is non stationary.

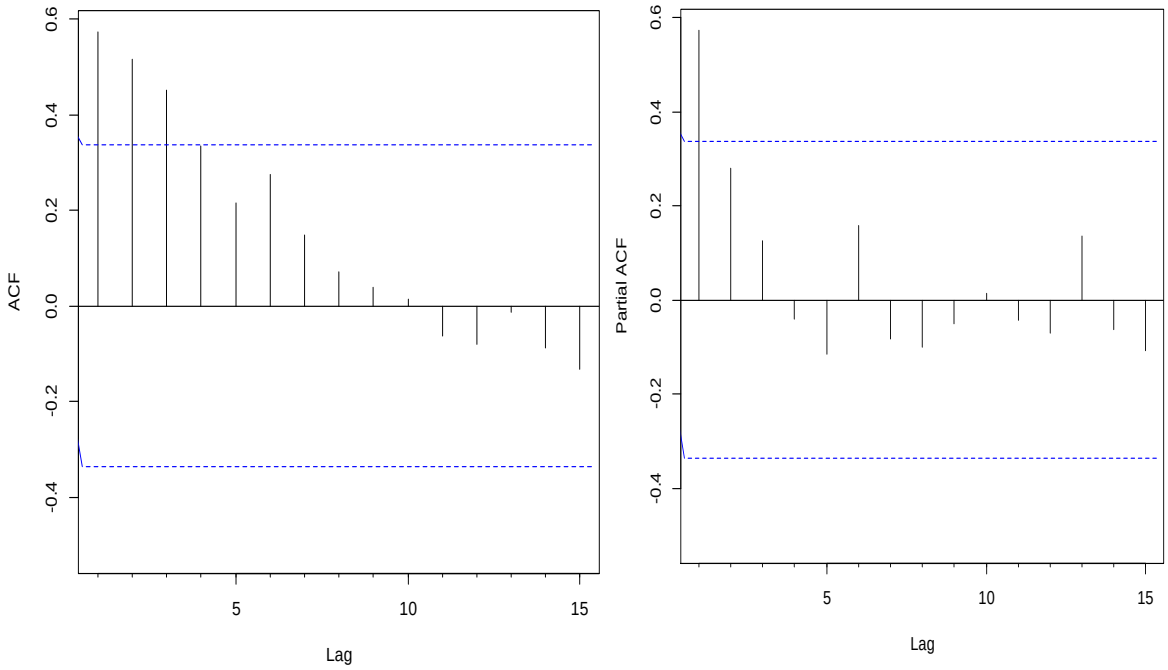
Augmented Dickey-Fuller Test

Dickey-Fuller = -2, Lag order = 3, p-value = 0.5

Thus, stationarizing the time series through differencing is needed so as to make the series stationary.

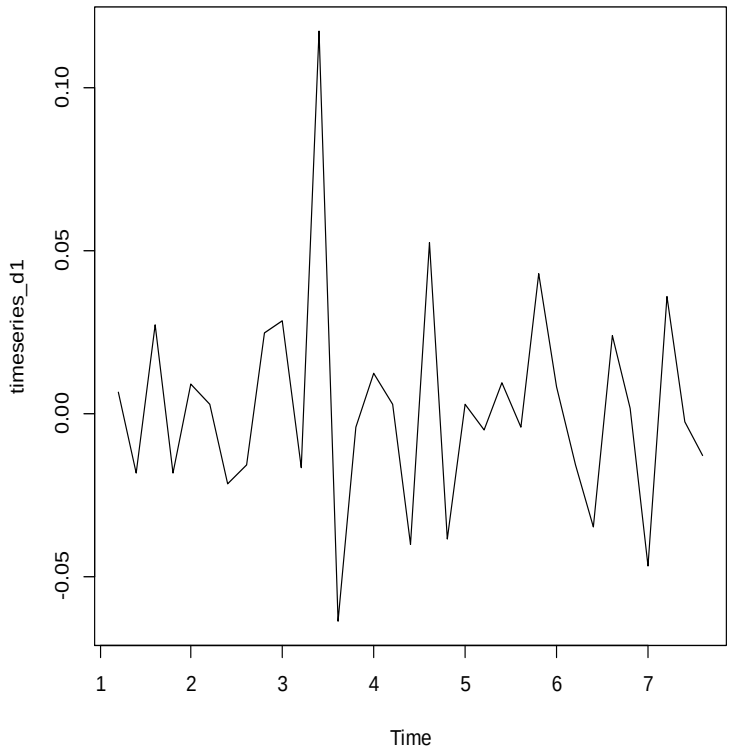
ACF and PACF Plots:

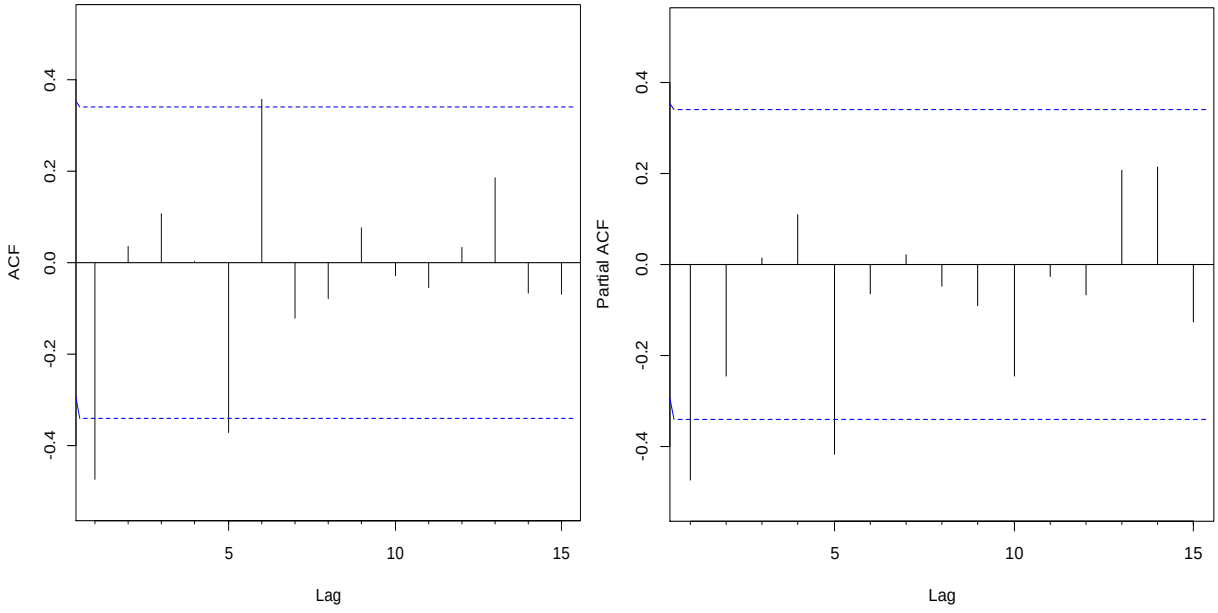
ACF plots display correlation between a series and its lags. In addition to suggesting the order of differencing, ACF plots can help in determining the order of the *M A (q)* model. Partial autocorrelation plots (PACF), as the name suggests, display correlation between a variable and its lags that is not explained by previous lags. PACF plots are useful when determining the order of the *AR(p)* model.



We demonstrate that the legs on the ACF chart have a downtrend and are set to zero. On the PACF chart, only the first leg is greater than the upper limit of confidence. Due to the lags above the upper line in ACF plot, we start difference-stationary process by d=1.

Now, the series is stationary and we can consider the changes in lags in two figures of ACF and PACF as below .Contrary to the ACF chart in which the process is interrupted after 1 delay, in PACF plot, the process tends to go toward zero. Therefore, the probable model, after the differentiation, is a moving average model with an order of 1, MA (1).





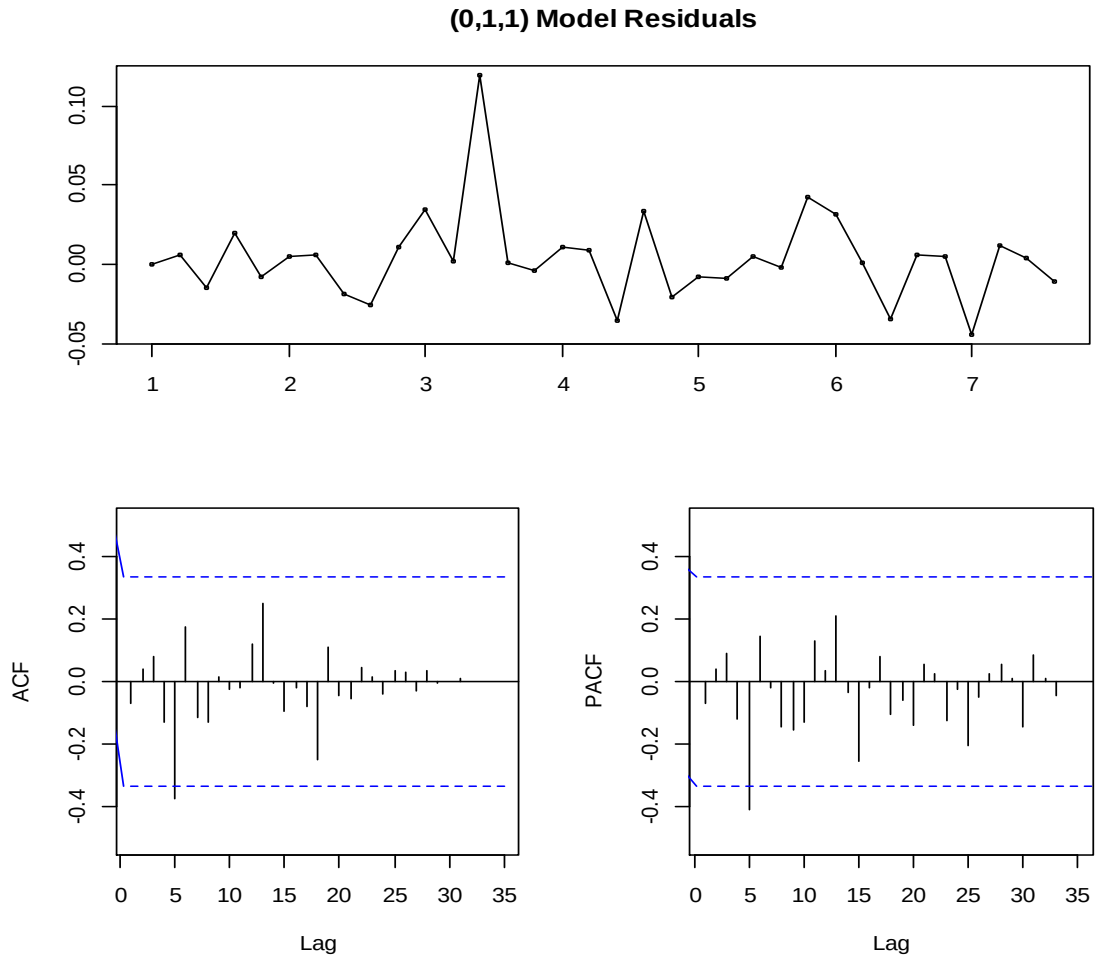
ARIMA MODEL:

The output for model ARIMA(0,1,1) is shown and model is written below:

sigma^2 estimated as 0.000829: log likelihood=70.6

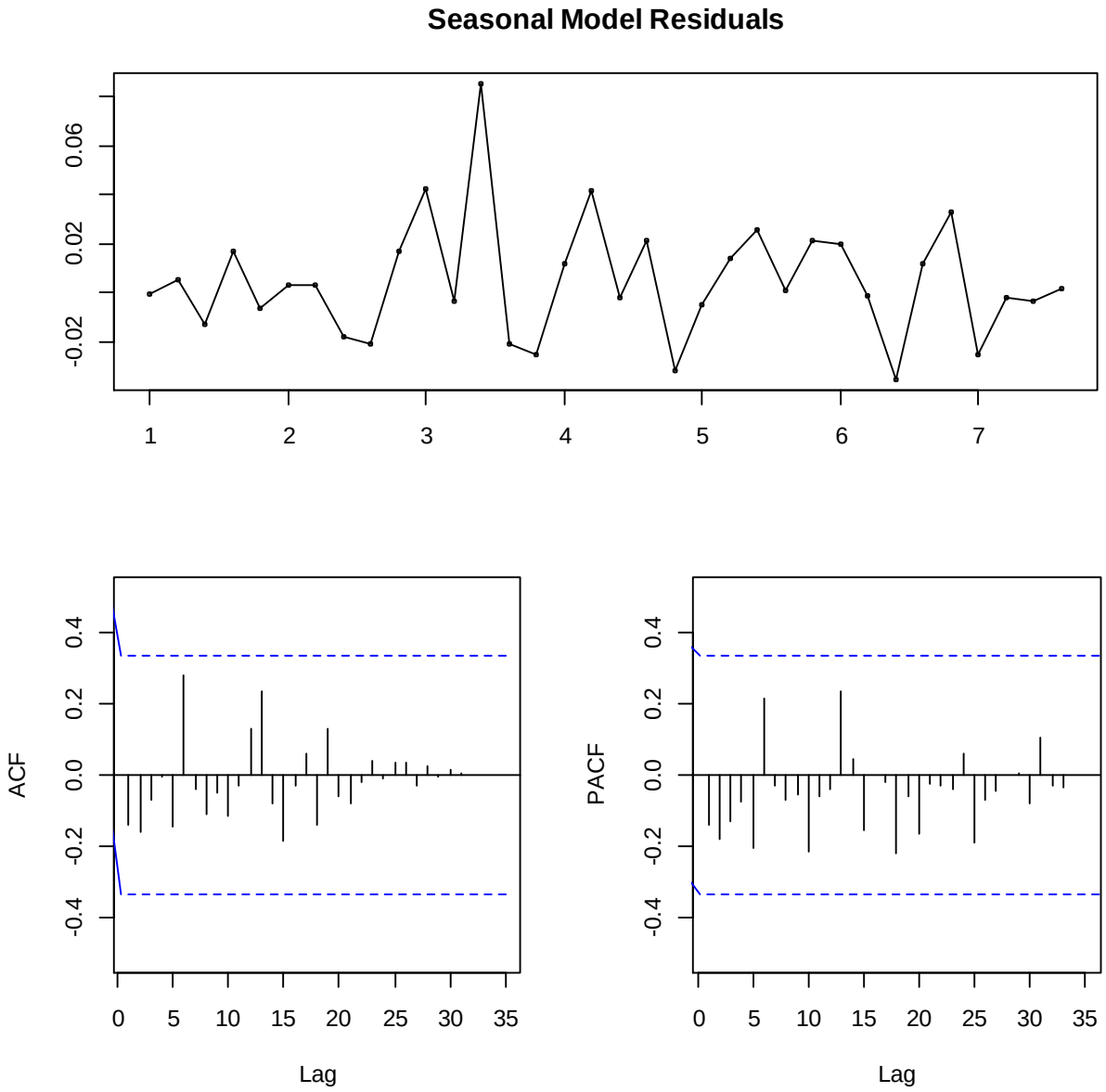
AIC= -137 AICc=-137 BIC=-134

$\hat{y}_{d_t} = -0.540e_{t-1} + E$



There is a clear pattern present in ACF/PACF and model residuals plots repeating at lag 5. This suggests that our model may be better off with a different specification, such as p = 5 or q = 5.

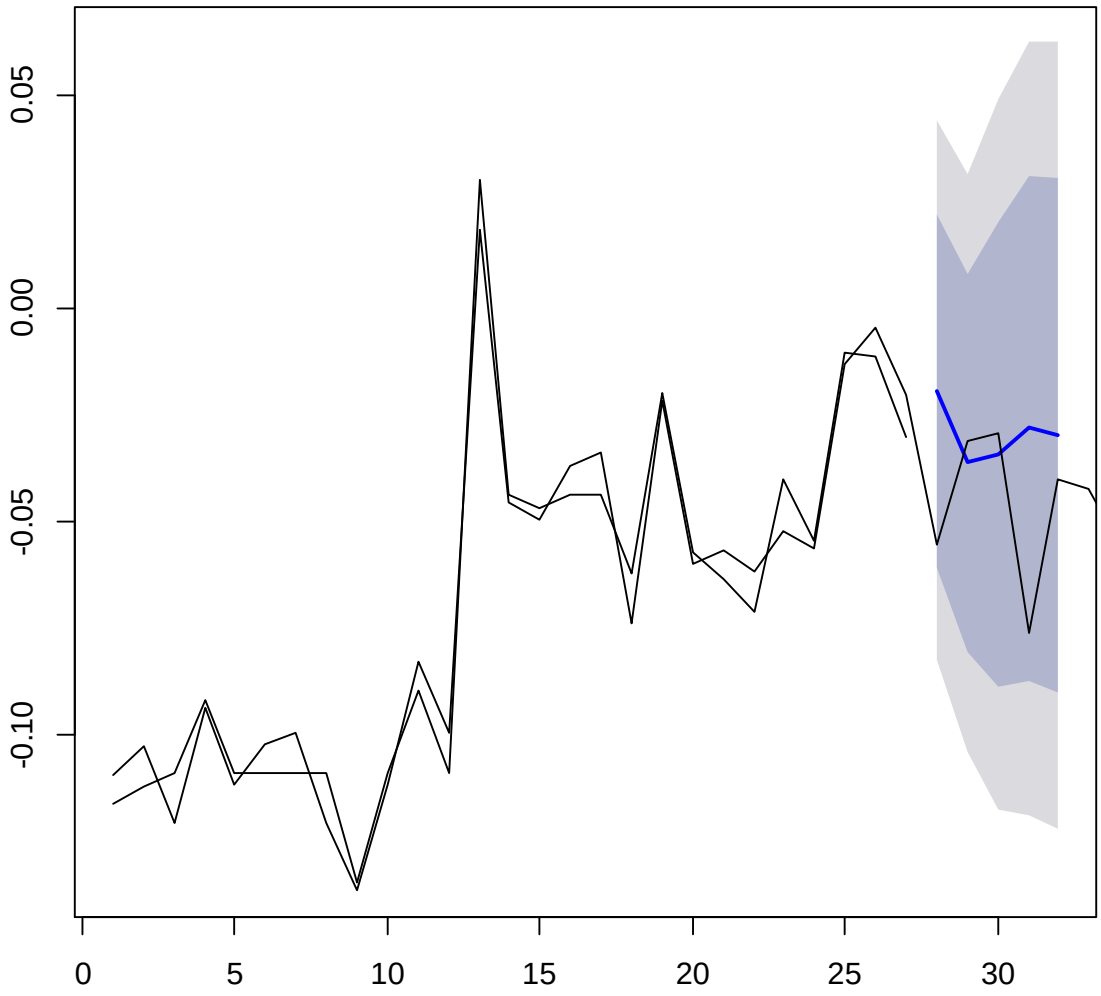
We fitted a new ARIMA model by orders (0,1,5) and it seems that we have a smaller error range, more or less centered around 0. We can observe that AIC is smaller as well.



sigma^2 estimated as 0.000599: log likelihood = 73.2, AIC = -134

Prediction:

We can specify forecast horizon *h* periods ahead for predictions to be made, and use the fitted model to generate those predictions:



Linear Regression is a linear approach to modelling the relationship between a scalar response or dependent variable and one or more explanatory variables or independent variables. (Wikipedia).

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-72.45706	10.60719	-6.83	1.2e-07	***
Data_1$year	0.03593	0.00526	6.83	1.2e-07	***
Data_1$month	-0.00217	0.00127	-1.70	0.098	.

Residual standard error: 0.0245 on 31 degrees of freedom

Multiple R-squared: 0.631, Adjusted R-squared: 0.607

F-statistic: 26.5 on 2 and 31 DF, p-value: 1.97e-07

As outlines above, model p-value is very small and it means that the model is appropriate. In addition, the variable of month doesn't affect Gross Domestic Product Growth Rate and should be out of the model. Therefore, the linear model fitted to the data is :

$$\hat{y}_i = -72.45706 + 0.03593 * year_i + E$$

Models Validation:

Cross validation for linear model :

folds	1	2	3	4	5
Predicted	-0.070471	-0.0726	-0.0748	-0.01504	-0.01721
Gdp(Real value)	-0.071300	-0.0403	-0.0544	-0.01050	-0.01140
CV residual	-0.000829	0.0323	0.0204	0.00454	0.00581

mse

0.000545

Cross validation for final model ARIMA(0,1,5) :

RMSE

[1] 0.0257

We applied cross validation method on presented models, root mean square error in ARIMA(0,1,5) equals 0.0257 and mean square error in linear model equals 0.000545 (root mean square error=0.02334) , hence both models work perfectly and the value of error in both are very close to each other though the linear model is slightly better.

Codes:

```
Data<-read.csv("C:\\Users\\mehr6\\Desktop\\time series\\final_data.csv", header = TRUE)
# split date column into three separate columns(month,day,year) and then merge them into the original data-frame
realdate<-as.Date(Data$date,format="%d-%m-%Y")
 dfdate <- data.frame(date=realdate)
year=as.numeric (format(realdate,"%Y"))
month=as.numeric (format(realdate,"%m"))
day=as.numeric (format(realdate,"%d"))

#merge them into the original data-frame
Data_1<-cbind(Data,day,month,year)
colnames(Data_1)
fix(Data_1)
library('ggplot2')
library('forecast')
library('tseries')
Data$date = as.Date(Data_1$date,format="%d-%m-%Y")
plot2 <- ggplot(Data_1, aes(date, gdp, group = 1)) +
        geom_point() +
        geom_line() +
        labs(x = "Date", y = "GDP",
            title = "Gross domestic product growth rate")
plot2
#analysis of seasonality, trend, and cycle (decomposing the data)
timeseries= ts(na.omit(Data_1$gdp), frequency=5)
decomp = stl(timeseries, s.window="periodic")
deseasonal_gdp <- seasadj(decomp)
```

```
plot(decomp)
#Stationarity and nonStationarity
adf.test(timeseries, alternative = "stationary") # The series is not stationary
#Autocorrelations
Acf(timeseries, main='')
Pacf(timeseries, main='')
#Differencing of order 1 terms
timeseries_d1 = diff(deseasonal_gdp, differences = 1)
plot(timeseries_d1)
adf.test(timeseries_d1, alternative = "stationary")
#ACF and PACF plots for differenced data
Acf(timeseries_d1, main='ACF for Differenced Series')
Pacf(timeseries_d1, main='PACF for Differenced Series')
#ARIMA Model
model<-auto.arima(deseasonal_gdp, seasonal=TRUE)
model
tsdisplay(residuals(model), lag.max=35, main='(0,1,1) Model Residuals')
#ARIMA Model_2
Model_2 = arima(deseasonal_gdp, order=c(0,1,5))
Model_2
tsdisplay(residuals(Model_2), lag.max=35, main='Seasonal Model Residuals')
#Forcasting the future data
ForC <- forecast(Model_2, h=5)
plot(ForC)
#Cross validation for final model (ARIMA(0,1,5)) [3 different ways!!]
train = Data_1[1:27,1]
test = Data_1[28:34, 1]
fit = Arima(train, order=c(0,1,5))
```

```
preds = as.vector(forecast(fit, h = length(test))$mean)
RMSE = sqrt(mean((preds - as.vector(test)) ^ 2))
RMSE
fcast <- forecast(fit,h=5)
plot(fcast, main=" ")
lines(ts(deseasonal_gdp))
#Linear regression
fitreg <- lm(Data_1$gdp ~  Data_1$year +Data_1$month  , data=Data_1)
summary(fitreg )
coefficients(fitreg )
#Cross Validation for Linear Regression
library('DAAG')
cv.lm(data=Data_1, fitreg, m=1)
```