

**STATISTICS FOR DATA ANALYTICS CONTINUOUS ASSESSMENT - REPORT**

**Multiple Regression and Time Series Analysis**

**Submitted by**

Kamesh Munuswamy - x19199562

MSc in Data Analytics - Batch A - 2020/21

# MULTIPLE REGRESSION MODEL

**Objective of the analysis:**

Motive of this analysis is to predict the female employee rate of 131 countries using Multiple Regression model.

**Data context:**

This metadata comprises four different datasets from distinct sectors browsed from UN data site. The idea behind the creation of this dataset is to predict the female employee rate by incorporating four different data sets together. These four datasets contain the female employee rate information for those who are above fifteen years of age across the globe. By identifying the factors, that contribute to predict the employee rate from different sectors of employment such as Agriculture, Industry, Own Business. These sectors add value to predict the female employment rate percentage across different areas.

**Data Source:**

All the four datasets are downloaded from the mentioned below link and combined to form a single metadata which is been used in this analysis. From the four different datasets, only the attributes which contributes for the employee rate for prediction is utilised. Data transformation of raw data is done using R Language and the model implementation is carried out in IBM SPSS Tool.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Summary of Data Source** | **Record Count** | **Link** |
| Female employee rate | Employee percentage rate | 2454 | <http://data.un.org/Data.aspx?d=GenderStat&f=inID%3a114#GenderStat> |
| Percent employed in agriculture | Employee percentage rate in agriculture | 3095 | <http://data.un.org/Data.aspx?d=GenderStat&f=inID%3a111> |
| Percent employed in industry | Employee percentage rate in Industry | 3095 | <http://data.un.org/Data.aspx?d=GenderStat&f=inID%3a112> |
| Percent employed in own account workers | Employee percentage rate in own account workers | 1812 | <http://data.un.org/Data.aspx?d=GenderStat&f=inID%3a116> |

**Variables Insight:**

In this dataset there are totally 3 predictors (X), each independent variable is taken from different of the sectors such as **Agriculture**, **Industry**, **Own account workers**. All these sectors hold the data of different countries which contributes in predicting the employee rate called **Value** which acts as a dependent variable (Y).



Nominal Measure → Country

Independent Variables → Agriculture (**X1),** Industry (**X2**), Own account workers (**X3**)

Dependent Variable → Employee rate (Y)

**Data Pre-processing and Transformation Process:**

Data has been pre-processed in R by merging the 4 different datasets together. Data transformation is carried out using different packages and the pre-processing steps has by been explained below.

1. Raw csv files were extracted from the source (UN data site) and using R language merging function was carried out by importing the data into R using read.csv function. The fundamental analysis was done when importing the file header and casting the data types.
2. Dependent and independent variables are merged together from unique datasets using **Inner join** function in **DPLYR** package. Inner join is carried out using the key attributes of **country** and **year** in the datasets.
3. In the process of data validation, **NULL** check was carried out in the newly created metadata. NULL values were cleansed using Drop function in **TDLYR** package.
4. At last, Metadata was exported has CSV file which is the input file for IBM SPSS for further modelling.

**Model Assumptions:**

* The dependent variable should be measured on a linear scale (i.e. is either an intervals or ratio)
* Data set should contain two or more independent variable which can be either continuous or categorical (i.e., an **ordinal** or **nominal** variable).
* Data should have **independence of observations (i.e. Independence of residuals), Durbin-Watson** statistics is used to check this constrain which is available in SPSS.
* Dependent and independent variables should possess linear relationship.
* Data needs to show **homoscedasticity.**
* Independent variables should not showcase **multicollinearity**. This can be interpreted using correlation coefficients and Tolerance/VIF values in SPSS statistics.
* There should be **no significant outliers**, **high leverage points** or **highly influential points** in the data.
* Check for **residual errors** in data which can be examined using histogram (with a superimposed normal curve) and a Normal P-P Plot in SPSS statistics.

**Implementation Process in IBM SPSS Statistics:**

* Pre-processed metadata CSV file from R is imported onto SPSS statistics tool to construct a multiple regression model and further analysis is interpreted on the predicted value.
* In SPSS statistics software following option is used to generate multiple regression model.

Analyse 🡪 Regression🡪Linear

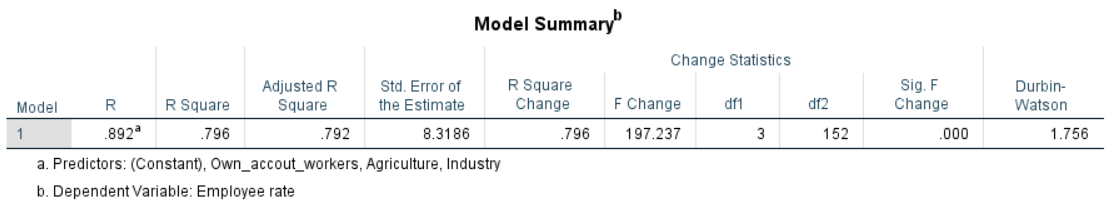
* Using the dependent and predicting variables following assumption related features are chosen in SPSS statistics for accurate analysis.
* Multiple regression model has to satisfy set of pre assumptions of data that are to be interpreted in the model. Based on the assumption following features are verified,

ⅰ- Durbin-Watson, ⅱ- Case-wise diagnostics, ⅲ- Covariance matrix, ⅳ- Confidence intervals, ⅴ- Model fit Collinearity diagnostics, ⅵ- Histogram and Normal probability plot, ⅶ- Mahalanobis, ⅷ- Cooks Value

**Result analysis and Interpretation:**

Based on the output generated by SPSS statistics for Multiple Regression model following interpretation is made based on the assumptions explained below,

### **Determining how well the model fits:**

****

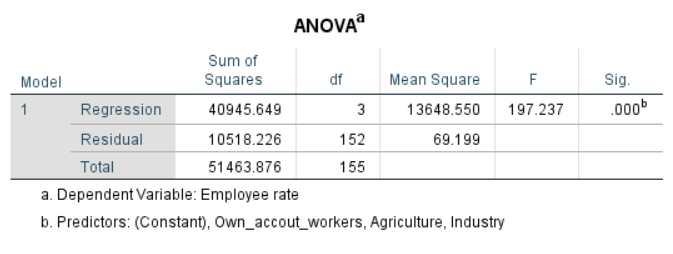
The column "R" reflects the value of R, the multiple correlation element. R can be regarded as a consistency indicator of the dependent variable's prediction; A value of 0.892 indicates a good prediction level in this case.

The "**R Square**" column represents the *R square* value (also called the coefficient of determination). Value of R square 0.796 determines that independent variable explains 79.6% of variability with respective to dependent variable (Employee rate).

"**Adjusted R Square**" (adj. R2) value .792 which is almost equal to R square value interprets that, all the independent variables are equally contributing in predicting the dependent variable. This value accurately reports the data used.

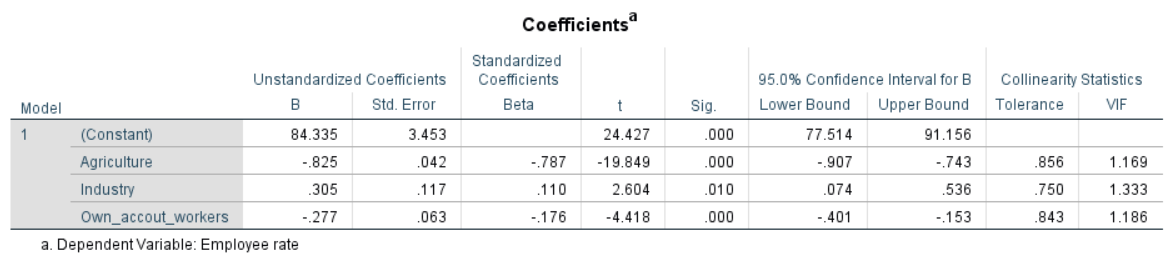
“**Durbin-Watson**” feature to check for autocorrelation between the variables gives the value of 1.756 through which it is assumed that data has few outliers and this smaller number of outliers will have no impact on our model. Which proves Assumption to absence of autocorrelation is not violated.

### **Statistical significance:**



**The F-ratio** (above mentioned table) in the **ANOVA** table checks whether the overall regression model is a good fit for the results. The table shows that the independent variables predict the dependent variable which is statistically significant, F (3, 152) = 197.237, p<.05 (i.e. the regression model is a strong data fit).

**Estimated model coefficients:**



The general equation to predict Employee rate (dependent variable) from Agriculture, Industry, Own account workers (independent variable) is,

Employee rate = 84.335 – (0.825 x agriculture) + (0.305 x Industry) – (0.277 x own\_account\_workers)

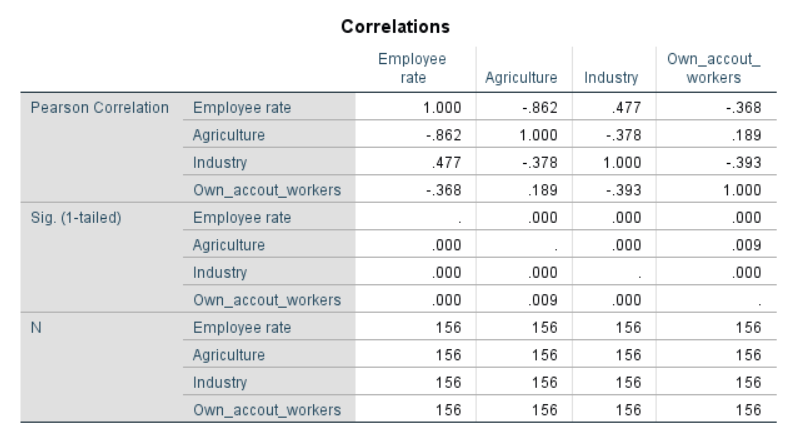
These values are obtained from the above coefficients table. Unstandardized coefficients reflect the amount of dependent variable differs when all other independent variables are kept constant.

**Statistical significance of the independent variables:**

Statistical significance of each independent variable is tested, if p > .05 it is considered that coefficients are statistically significant. From the “Sig” column its is clear that all the independent variables are statistically significant.

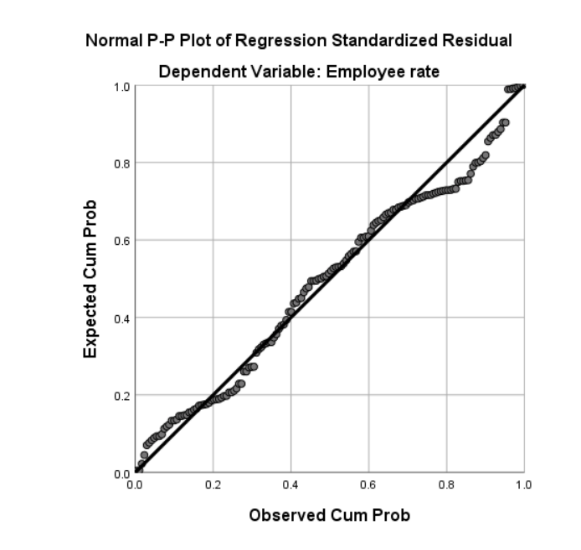
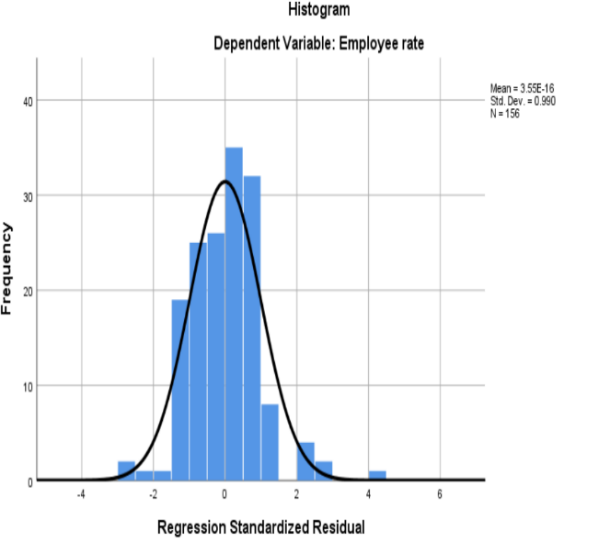
“**Variance Inflation factor (VIF)**” feature to test multi correlation between the predictor variables showcase values less than 10 which proves there is no multicollinearity among the independent variables.

**Correlation analysis:**



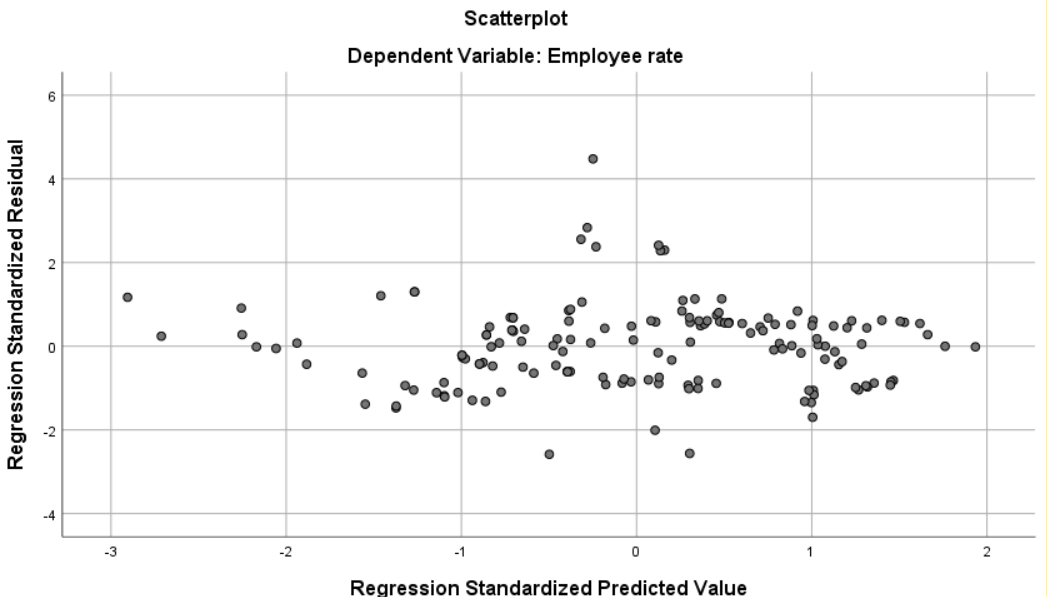
The above correlation chart gives the idea that there is a negative correlation prevails among the dependent and the independent variable. From the chart values it is clear that all the variables are not highly correlated.

**Histogram and P-P plot:**



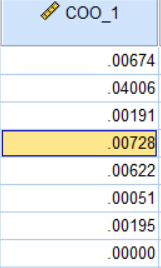
To check for the linearity assumption Histogram and P-P Plots are plotted. From the above charts it is satisfying to see that data is normally distributed with no great deviation and outliers in the data.

**Scatter Plot:**



Form the above scatter plot, assumption to check the **homoscedasticity** is verified. As the points are lies close to 0 line, which is scattered between -3 to 3. This proves **homoscedasticity assumption is not violated.**

The Cook value gives feedback that the model has no influential points as all the values are less than 1.



**Conclusion:**

A multiple regression was run to predict Employee rate from Agriculture, Industry, Own account workers. These variables statistically significantly predicted Employee rate, *F* (3, 152) = 197.237, *p* < .05, R square = .792. All three variables added statistically significantly to the prediction, *p* < .05.

**TIME SERIES – ARIMA MODEL**

**Definition:**

A time series is a sequence of normal, time-ordered measurements of a quantitative feature of an entity or collective phenomena taken at successive periods / points of time, in most cases equidistant.

**Objective of the analysis:**

The idea is to understand the time series models and their forecasting techniques for predicting the future data. In this analysis, ARIMA model technique is been implemented to forecast the total ICP (Installation Control Point) for the next 4 years of time using the past data.

**Data context:**

For this time series forecasting model, the dataset (raw data) which contains supply of oil data for all the European countries from the year 2016 to 2019 with monthly frequency is been used. We are considering only the Ireland country data for this analysis. Using this raw data of around 48 records, the aim motive is to forecast the supply of oil for the next 4 years of Ireland.

**Data Source:**

The dataset with supply of oil information of Ireland is sourced from Eurostat metadata -Energy statistics of supply, transformation and consumption. The raw data is extracted from the below mentioned link. <https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_102m&lang=en>

**Data Transformation:**

The original raw data is loaded into R during the data transformation process for the analysis to be carried out. The following steps to be performed are as follows:

* Using the ts() function, the original raw data is translated into time series data and the related graphs, summary, start, end and frequency values are listed and tested to get the correct values.
* The associated mean line and seasonality are known by plotting a line map using the abline() function and using boxplots, separately.

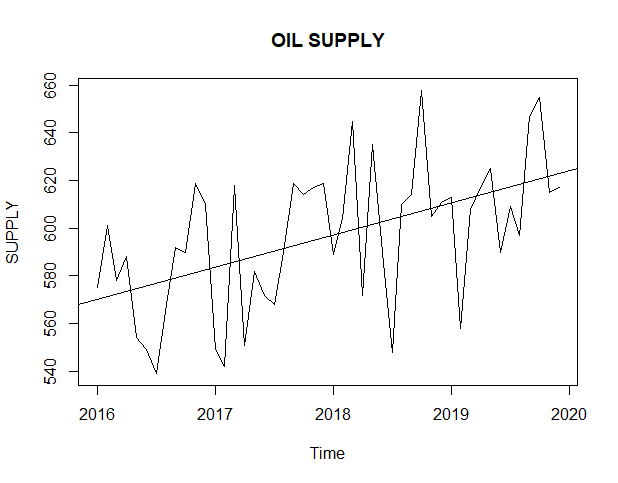
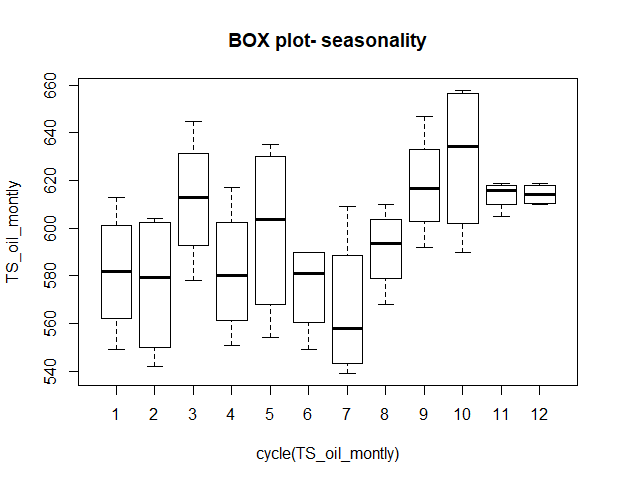
** **

Fig1.1 Fig1.2

**ARIMA model:**

An ARIMA model is a family of statistical models for time series data processing and forecasting.

AR: **Auto-regression**. A model that uses the relationship of dependency between an observation and a number of lagged observations.

I: **Integrated**. The use of differentiating raw observations (e.g., subtracting an observation from an observation during the previous time step) to stationary the time series.

MA: **Moving Average**. A model using the correlation between an observation and a residual error applied to lagged measurements.

**Classification of ARIMA model assumptions:**

Based on the observations of data using the above-mentioned graphs(fig1.1, fig1.2), it is observed that the data is non-stationary. Since it is mandatory for the data to be stationary in ARIMA model in order to build an accurate model for forecasting. The following measures are taken to convert non- stationary data into stationary data.

**Conditions for Stationary data:**

* The data should have a constant mean for the series of data it possesses.
* The time series data should have constant variance and covariance in time.

The most common approach to make the data with constant mean is by taking differentiation.

Depending on the complexity of the data series more than one differencing is required, which defines the d value for the model. If the time series is already stationary, then d = 0. To achieve constant variance and co variance, log transformation of power 10 is applied to the time series.

After applying the differentiation and log transformation to our data, underlying result showcase the stationarity of our time series data.

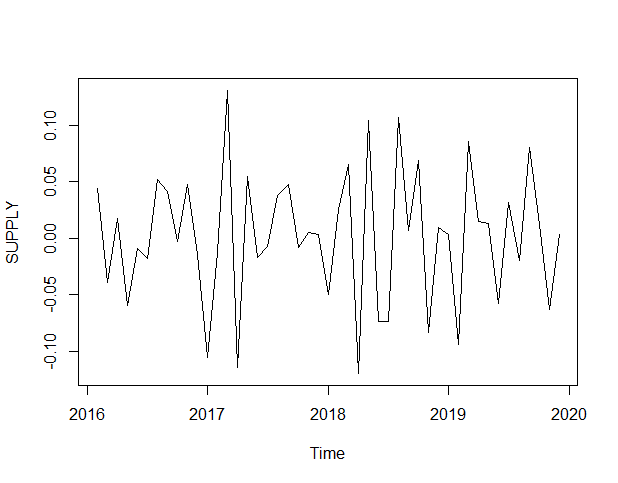
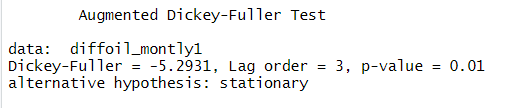


Fig1.3

**Augmented Dickey-Fuller Test:**

The Augmented Dickey Fuller Test (ADF) is unit root test for stationarity check in the time series data. This test is applied to our transformed data to check whether the data is stationary or non-stationary. In R adf.test() function performs this test, following result is displayed below.



ADF test demonstrate that for the data to be stationary its corresponding p-value should be less than .05 (p-value > 0.05). Considering the fact, we have p-value = 0.01 which confirms that our data after transformation is stationary.

**Fitting of model (ARIMA):**

Foremost step in building an ARIMA model is to define the parameters of ARIMA model.

A standard notation of the model is ARIMA (p, d, q). Components of ARIMA model defines,

* p → number of lag order in AR
* d → degree of differencing
* q → order of moving average in MA

**ACF and PACF Values:**

Using Autocorrelation function (ACF) and Partial Autocorrelation (PACF) plots, parameters of ARIMA model p and q is defined. In R, acf() and pacf() functions are used for plotting. ACF plot is merely a graph with coefficient of correlation between a time series and its own lag. Whereas, PACF plot is a plot of the *partial* correlation coefficients between the series and its lag itself.

Following acf() and pacf() plots are plotted to define the p and q values,

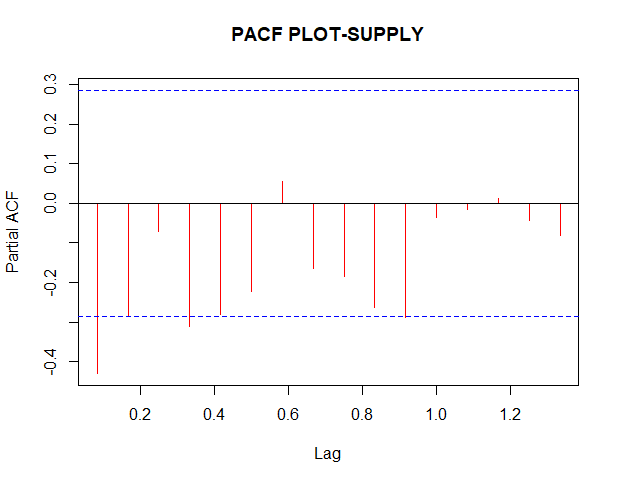
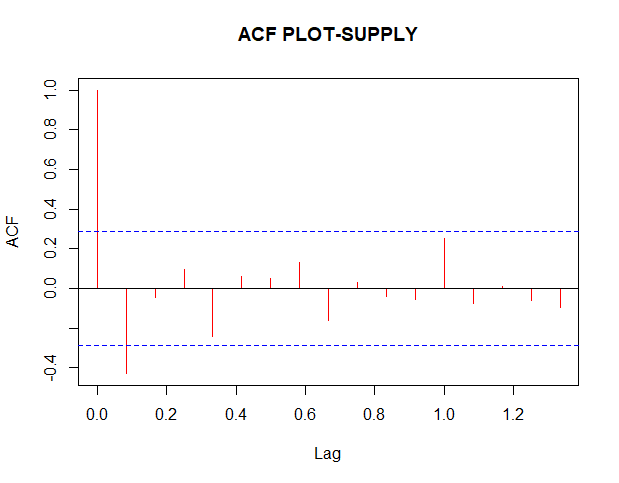
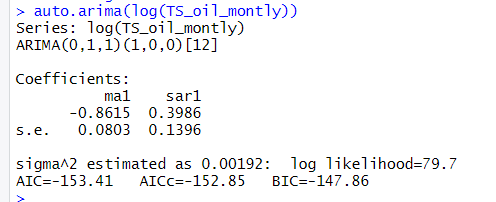


Fig1.4 Fig1.5

From the above figures, it is obvious to understand there is a sample ACF cut-off after 2 lag and the sample PACF cut off after 1 lag. So, we assume our p and q value is ARMA(0,1). Apparently, for the original time series, model fit should be ARIMA(0,1,1).

In order to find the best model fit for ARIMA, R provides a special function called auto.arima(), which provides the best fit model based on the least AIC value among the models. The function auto.arima() gives the values of seasonality, such as (P, D, Q)m as well where (P, D, Q) represents the seasonal auto-regressive component. The output of auto.arima() is displayed below.



We see that auto.arima() function returns an identical model ARIMA(0,1,1) as same as our predicted model. This confirms the best fit model as MA(1) along with the seasonal component

ARIMA(0,1,1)(1,0,0)[12] which can be termed as SARIMA model. It has the AIC value of 153.41 considered to be the least value when evaluated with other models.

**Forecasting values:**

Using the ARIMA model components and parameters values, we are predicting the values for next 4 years. The forecasted values of 4 years data along with graph representing the data series flow of future is displayed below (Fig1.6),

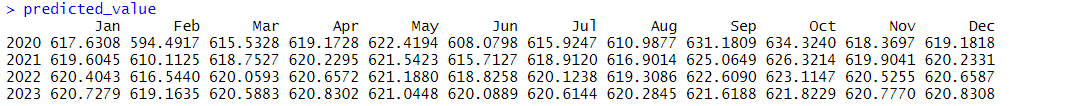


Fig1.6

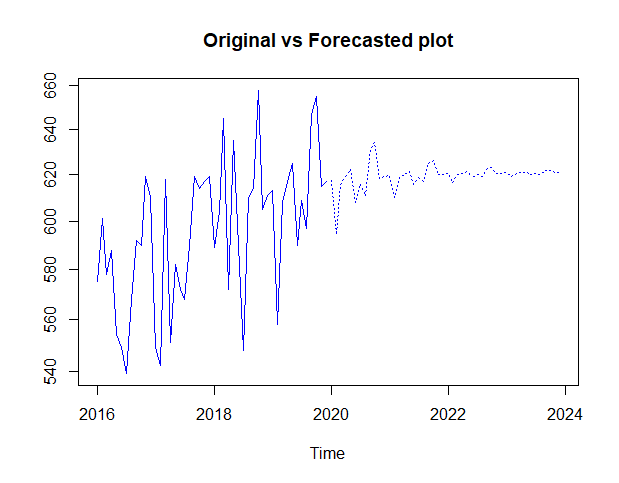
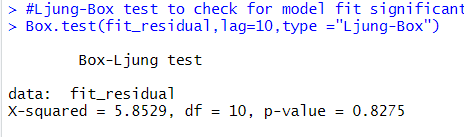


Fig1.7

**Result Evaluation:**

ARIMA model is tested and evaluated for its forecasted values using few techniques. Likewise, one of the test is ***Box- Ljung test*** to check for auto-correlation. From this test it is clear that p-value = 0.8275 which is greater than p-value = 0.05. Which confirms there is no auto-correlation exist.



Below plots (Fig1.8) explains about the ***Residuals*** wherein, it is clearly picturized that data is normally distributed with constant mean and variance. Proving the matter of fact that residuals are independent by showing normal distribution.

From the ***Summary()*** function, considering the important factors like AIC = -153.41 and RMSE= 0.04242223 and MAE= 0.0349741 followed by other errors determining factors satisfy that ARIMA(0,1,1) is a best fit model in this forecasting.

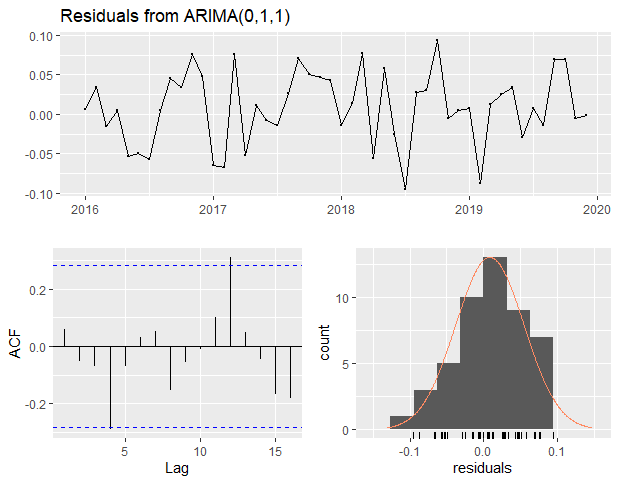
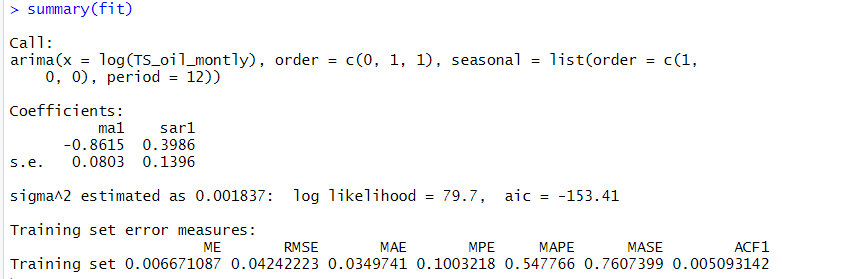
 

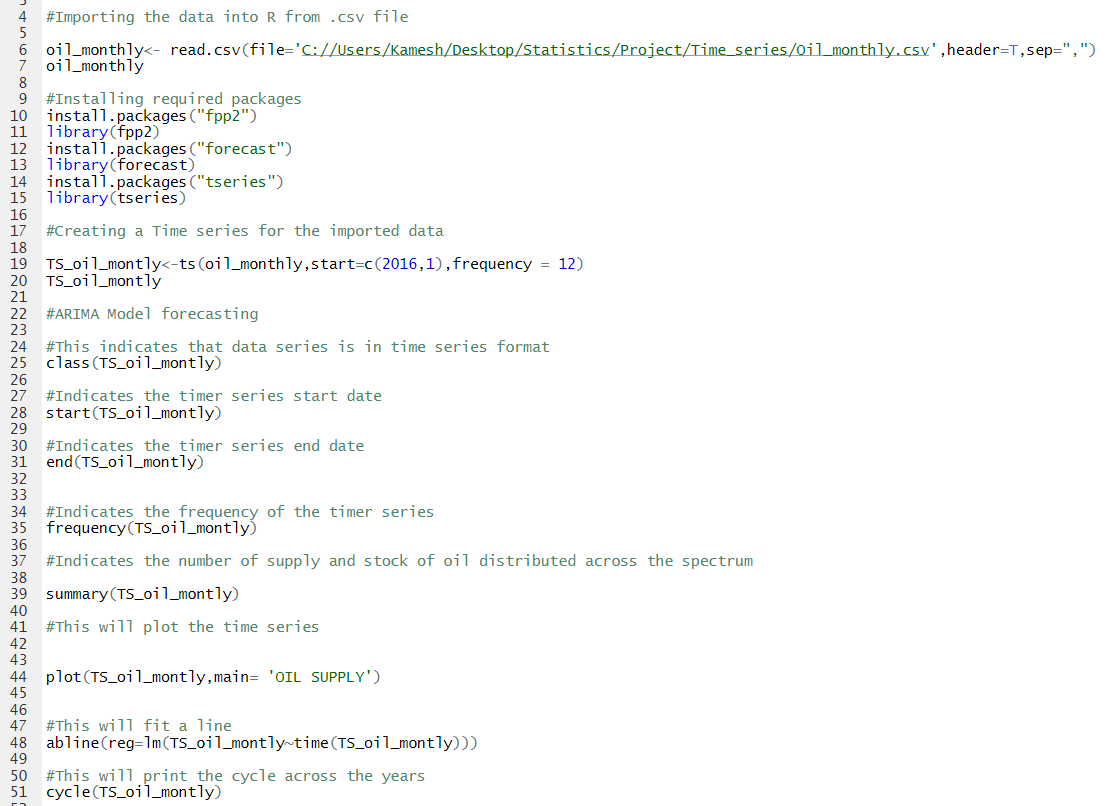
Fig1.8 Fig1.9

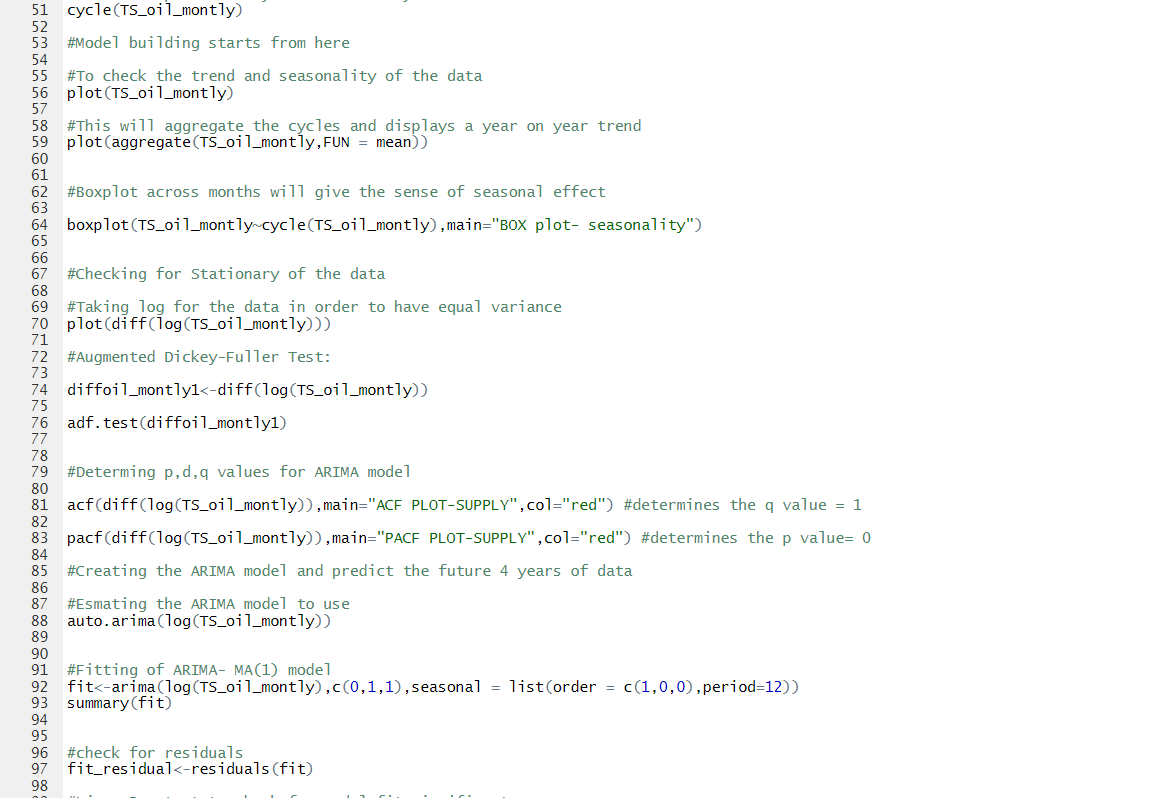
**Conclusion:**

To summarize, ARIMA model is built to forecast the oil supply of Ireland country. From the results obtained by ARIMA(1,1,0) model and through various evaluation and comparison of predicted value with actual value, the obtained accuracy of prediction is satisfying. Thus concludes, this model is working fine in forecasting.

**APPENDIX:**

R scripts snippet of time series to build ARIMA model.





**References:**

Levine, David M., Stephan, David., Szabat, Kathryn A. (2016) *Business statistics: a first course. 7th edn.* London: Pearson.

Box, George E. P., Jenkins, Gwilym M., Reinsel, Gregory C. (2008)*Time series analysis: forecasting and control.* 4th edn. Oxford: Wiley-Blackwell.

Riaz khan (2017) *ARIMA model for forecasting– Example in R.* Available at: <https://rpubs.com/riazakhan94/arima_with_example> [Accessed 3 March 2020].