National College of Ireland

STATISTICS FOR DATA ANALYTICS

ON

MULTIPLE REGRESSION

AND

TIME SERIES ANALYSIS

Sai Chethan Singu

Register Number: x18181937

Module: Statistics for Data Analytics

Date: 05-04-2020

Word count : 1955

(1) Hierarchical Multiple Regression

In supervised learning, there are three types of regression models namely standard, hierarchical and stepwise regression. Hierarchical regression which is also known as sequential regression is used for this model to study the independent variables which are predominantly influencing the dependent variable. Independent variables or a set of independent variables are entered in blocks or steps so that we can analyze the contribution of the individual independent variables with relation to the prediction of the dependent variable.

DATA SET USED:

The causes of deaths caused due to various diseases and factors like Cancer, Suicide, Accidents, pneumonia, Chronic diseases, Nervous disorders have been considered and extracted by combining various statistical tables from Eurostat website. The initial dataset comprises of 115 records. Data cleaning is done by removing the null values and the countries with empty values were removed. The resultant dataset consists of 106 records.

https://ec.europa.eu/eurostat/web/health/data/main-tables

Independent Variables	Dependent Variables
Cancer	All causes of deaths
Suicide	
Accidents	
Chronic diseases	
Nervous disorders	

OBJECTIVE OF THE ANALYSIS:

The main objective is to carry out the hierarchical multiple regression model to predict the mortality pattern of deaths (both males and females) by using various influencing factors.

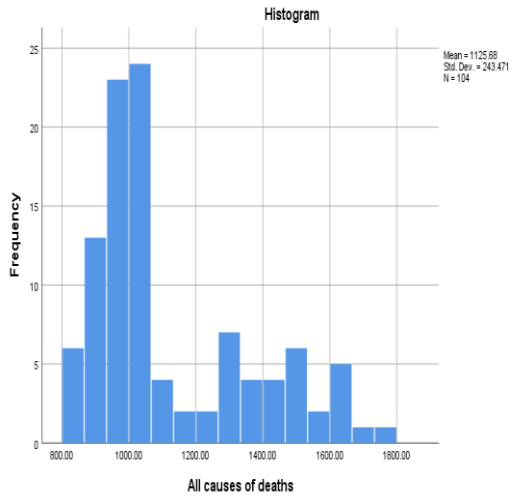
VERIFICATION OF ASSUMPTIONS:

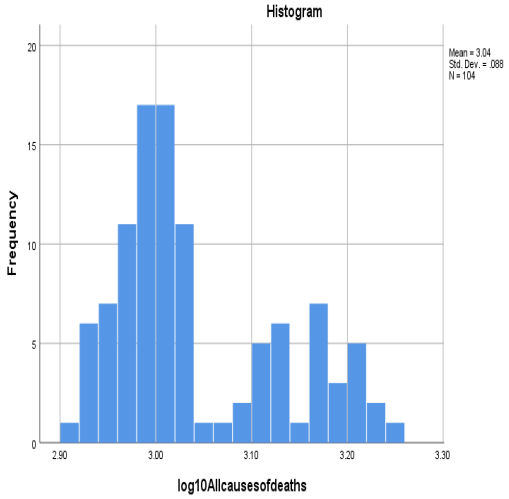
SAMPLE SIZE:

The sample size is calculated by using the formula N > 50+8M where 'M' Is the number of independent variables in the dataset. According to the formula, the value of N for this dataset is 90 and the sample size is 104. Hence assumption is satisfied.

NORMALITY TEST FOR THE DEPENDENT VARIABLE:

The independent variable "All causes of deaths" is plotted on the histogram plot. In which skewness and kurtosis were determined. The first histogram plot was left-skewed so in order to reduce the skewness, log to the base 10 transformation is used. Another histogram is plotted for log values where less skewness can be observed.





CORRELATIONS & MULTICOLLINEARITY:

The ideal correlation between dependent and independent variables should be close to one. For this, we can check the correlation table. From the below table we can see the correlation the dependant variable (All causes of deaths) and independent variables (cancer, suicide, accidents, chronic diseases, nervous disorders) are 0.55, 0.45, 0.39, 0.91 and -0.60 respectively. We can see the correlation between them is high.

Correlations

		All causes of deaths	Cancer	Suicide	Accidents	Chronic diseases	Nervous disorders
Pearson Correlation	All causes of deaths	1.000	.554	.459	.397	.913	-.607
	Cancer	.554	1.000	.594	.449	.636	-.391
	Suicide	.459	.594	1.000	.684	.549	-.261
	Accidents	.397	.449	.684	1.000	.471	-.271
	Chronic diseases	.913	.636	.549	.471	1.000	-.645
	Nervous disorders	-.607	-.391	-.261	-.271	-.645	1.000
Sig. (1-tailed)	All causes of deaths	.	.000	.000	.000	.000	.000
	Cancer	.000	.	.000	.000	.000	.000
	Suicide	.000	.000	.	.000	.000	.004
	Accidents	.000	.000	.000	.	.000	.003
	Chronic diseases	.000	.000	.000	.000	.	.000
	Nervous disorders	.000	.000	.004	.003	.000	.
N	All causes of deaths	104	104	104	104	104	104
	Cancer	104	104	104	104	104	104
	Suicide	104	104	104	104	104	104
	Accidents	104	104	104	104	104	104
	Chronic diseases	104	104	104	104	104	104
	Nervous disorders	104	104	104	104	104	104

In order to check multicollinearity, VIF and Tolerance factors are taken into consideration. It can be seen from the below table. The ideal value for Tolerance needs to be higher than 0.10. All the independent variables in this data are greater than the specified value. The variance

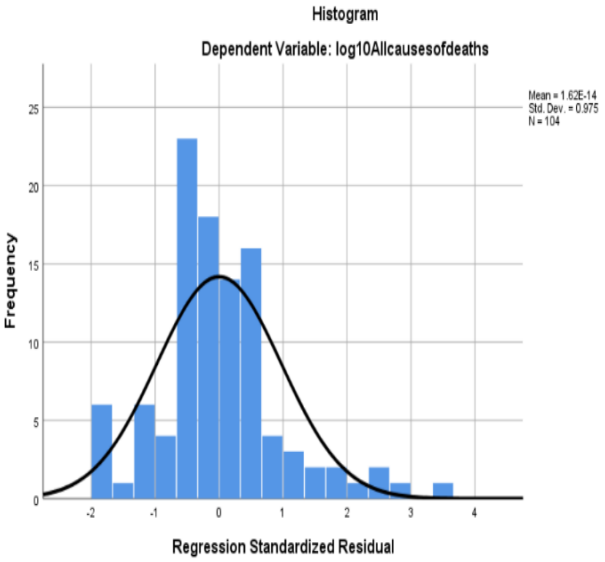
inflation factor (VIF) should not exceed 10. From the table, VIF values for cancer, suicide, accidents, chronic diseases and nervous diseases are 1.7, 1.9, 2.4, 2.7 < 10. Therefore correlation and collinearity assumptions were satisfied.

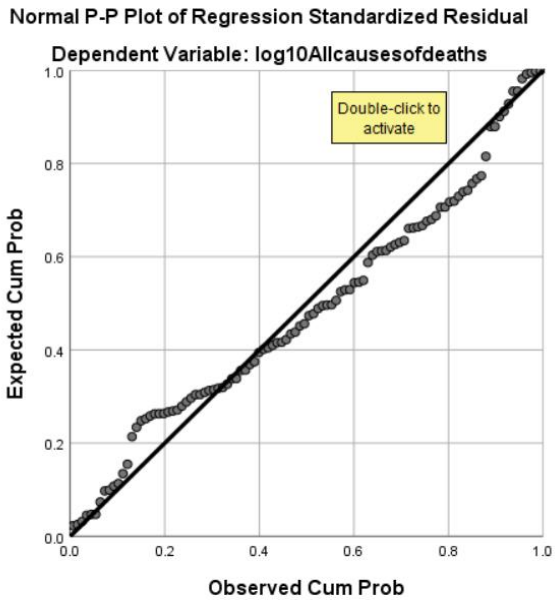
Coefficients[a]

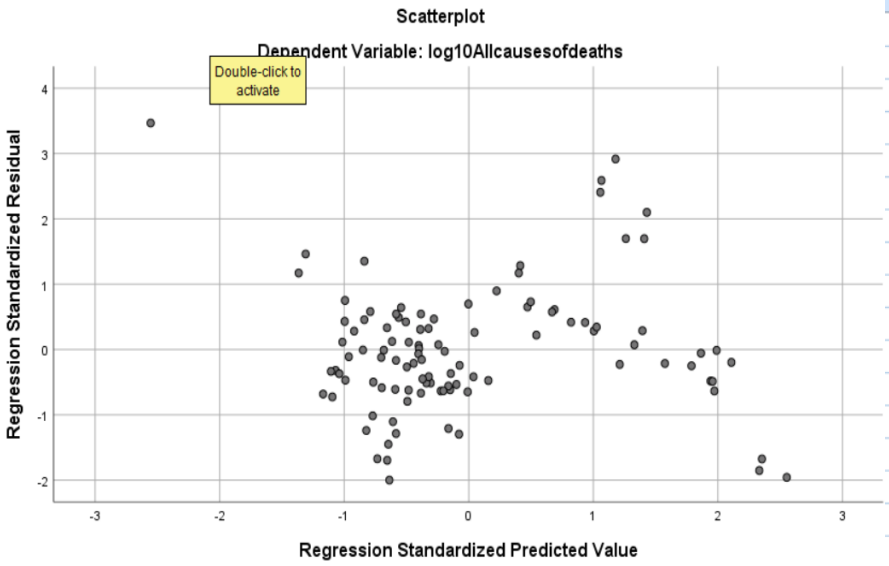
Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	217.609	157.692		1.380	.171	-95.211	530.428					
	Cancer	2.960	.688	.434	4.300	.000	1.595	4.325	.554	.393	.349	.647	1.545
	Suicide	9.227	4.627	.201	1.994	.049	.048	18.406	.459	.195	.162	.647	1.545
2	(Constant)	622.288	99.421		6.259	.000	424.990	819.587					
	Cancer	-.170	.390	-.025	-.437	.663	-.944	.604	.554	-.044	-.018	.509	1.965
	Suicide	-1.759	2.922	-.038	-.602	.549	-7.558	4.039	.459	-.061	-.025	.410	2.439
	Accidents	-.326	1.200	-.015	-.272	.786	-2.707	2.054	.397	-.027	-.011	.518	1.932
	Chronic diseases	4.406	.315	.942	14.008	.000	3.782	5.030	.913	.817	.572	.369	2.712
	Nervous disorders	-.365	.847	-.023	-.430	.668	-2.045	1.316	-.607	-.043	-.018	.571	1.752

a. Dependent Variable: All causes of deaths

HOMOSCEDASTICITY, RESIDUAL'S INDEPENDENCE AND OUTLIERS:







MAH_1
20.27729
19.63855
19.04909
14.34340
14.28207
12.03141
11.71973
11.48942
11.44935
11.34544
10.29103
10.21572
8.52357
8.38853
8.11699
7.92563
7.90200

From the above graphs, standardized residual is normally distributed and the probability plot also follows linearity. While coming to the scatter plot, it is also rectangularly distributed. By neglecting few outliers, Standardized residual value is in the range of +3.3 to -3.3. Outliers can be checked by using Mahalanobis distance (MAH_1) which will be generated by regression analysis. As per the Residual Statistics table for five independent variables the critical value is 20.52. In this data, Mahalanobis distance is within the acceptable rate.

EVALUATION OF THE MODEL:

By performing Hierarchical multiple regression analysis on the model, various factors are evaluated. As we have used hierarchical regression two independent variables were selected in the first model and the remaining four independent variables are selected in the second model. The first R value (.577) is the correlation between the dependent variable and independent variables which includes suicide and cancer. It accounts for 33% variability in all causes of deaths. In the second model by adding other independent variables namely nervous disorders, accidents and chronic diseases the value of R is (.915) and it accounts for 83% variability in predicting dependent variable. For both the models the adjusted R square is very close to the value of R square which means that they are equally contributing in predicting dependent variable.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.577[a]	.333	.320	200.78314	.333	25.226	2	101	.000
2	.915[b]	.837	.828	100.91532	.503	100.606	3	98	.000

a. Predictors: (Constant), Suicide, Cancer

b. Predictors: (Constant), Suicide, Cancer, Nervous disorders, Accidents, Chronic diseases

c. Dependent Variable: All causes of deaths

By seeing the table below (ANOVA) statistical significance can be found out. From the overview of the table for both the models after performing hierarchical multiple regression, all the sigma values are statistically significant ($p < 0.0005$).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2033926.690	2	1016963.345	25.226	.000[b]
	Residual	4071700.648	101	40313.868		
	Total	6105627.338	103			
2	Regression	5107604.887	5	1021520.977	100.307	.000[c]
	Residual	998022.451	98	10183.903		
	Total	6105627.338	103			

a. Dependent Variable: All causes of deaths

b. Predictors: (Constant), Suicide, Cancer

c. Predictors: (Constant), Suicide, Cancer, Nervous disorders, Accidents, Chronic diseases

EVALUATION OF INDEPENDENT VARIABLES:

In the coefficients table, the largest beta coefficient is 0.94 for chronic diseases which means that this independent variable (chronic diseases) uniquely contributes to the prediction of the dependent variable. And low beta (0.15) value by accidents means that it has less unique contribution with respect to the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	217.609	157.692		1.380	.171	-95.211	530.428					
	Cancer	2.960	.688	.434	4.300	.000	1.595	4.325	.554	.393	.349	.647	1.545
	Suicide	9.227	4.627	.201	1.994	.049	.048	18.406	.459	.195	.162	.647	1.545
2	(Constant)	622.288	99.421		6.259	.000	424.990	819.587					
	Cancer	-.170	.390	-.025	-.437	.663	-.944	.604	.554	-.044	-.018	.509	1.965
	Suicide	-1.759	2.922	-.038	-.602	.549	-7.558	4.039	.459	-.061	-.025	.410	2.439
	Accidents	-.326	1.200	-.015	-.272	.786	-2.707	2.054	.397	-.027	-.011	.518	1.932
	Chronic diseases	4.406	.315	.942	14.008	.000	3.782	5.030	.913	.817	.572	.369	2.712
	Nervous disorders	-.365	.847	-.023	-.430	.668	-2.045	1.316	-.607	-.043	-.018	.571	1.752

a. Dependent Variable: All causes of deaths

CONCLUSION:

The Hierarchical multiple regression is used to predict the total causes of deaths by using different independent variables (cancer, suicide, accidents, chronic diseases and nervous diseases). Different analyses were conducted to make sure the assumptions of linearity, multicollinearity, normality and homoscedasticity are satisfied. In model 1 explaining F change (2,101) = 25.22 and 33% of variance in predicting dependent variable (All causes of deaths) and sigma is less than 0.005. When compared to model 2, it is contributing 83% of variance in prediction and F change (3,98) = 100.60.

(2) TIME SERIES ANALYSIS

It is the orderly sequence of values of a certain variable at regular intervals of time. Drawing the patterns by using the information of timed intervals to predict future behaviour is defined as time series analysis. The different components of time series analysis include Trend, Seasonality, irregularity and noise. By analyzing trend and seasonality, we can predict or forecast the future. Normally trend describes the gradual increase or decrease over a time period. Fixed periodical fluctuations in the time series is called as seasonality. Irregularity or random variations in the time series is due to out of control or unpredictable movements.

DATA SET USED:

The dataset used is unemployment percentage of labour force for those aged between 15 to 74. It consists of Forty countries quarterly data. For doing time series analysis, in those countries, Sweden is selected. There are no missing values present for Sweden country. From 2009 quarter 1 to 2019 quarter 4 data is considered for performing analysis that will count to 44 records.

https://ec.europa.eu/eurostat/databrowser/view/tipsun30/default/table?lang=en

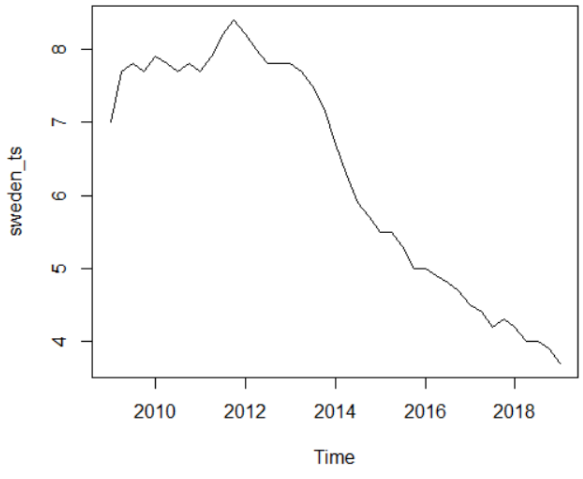
OBJECTIVE OF THE ANALYSIS:

The objective is to perform time series analysis to predict the future values (Upcoming years) for unemployment by analysing various components like trend, seasonality and others.

ANALYZING THE DATA:

The country named Sweden data is taken in an excel file (test1) which comprises 44 records. The excel file is read by using 'readxl' library. Then by using timeseries function, the data is represented as a time series model. The data is taken from the year 2009 to 2019 that is indicated by start and end functions. Frequency= 4 denotes that data is distributed quarterly. By plotting the timeseries data, a downward trend with some irregularities is clearly observed.

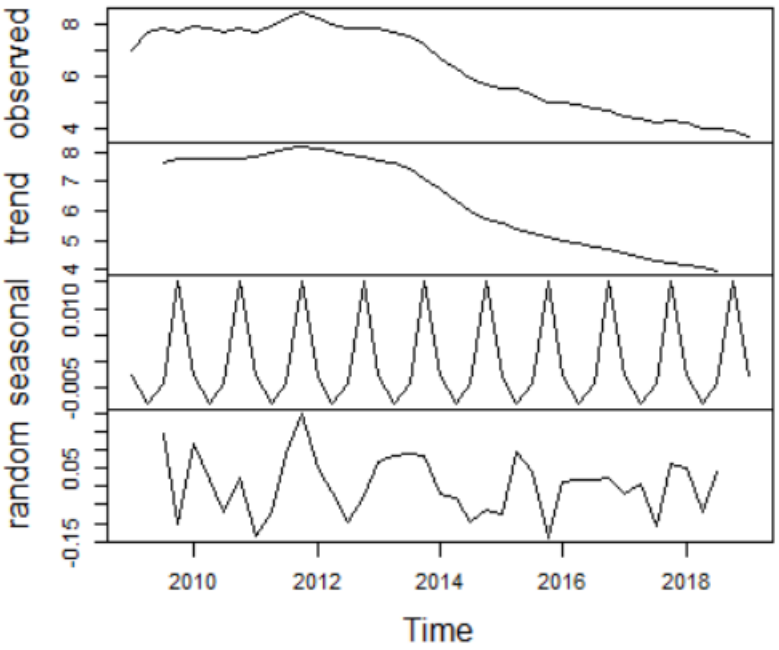
```
#Reading and plotting data in timeseries
data= read_excel("C:/Users/saich/Desktop/test1.xlsx")
data
sweden_ts=ts(data[,2],start=2009,end=2019,frequency = 4)
plot(sweden_ts)
```



VERIFYING DATA IS STATIONARY OR NON-STATIONARY:

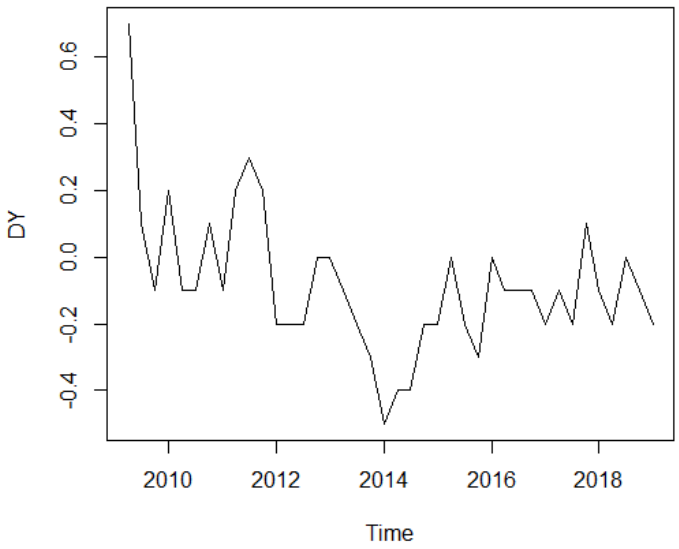
Data can be checked whether it is stationary or non-stationary by using decompose function, where we can see the trends, seasonality and randomness in the data. From the below graph, we can observe, there is a downward linear trend and there is no seasonality in the data as there are no continuous cyclic patterns. We can also see the trend and seasonality values by printing the decompose function. So from the graph, we can conclude that graph is not stationary.

```
#Decomposing data
tsdecompose=decompose(sweden_ts)
plot(tsdecompose)
tsdecompose
```



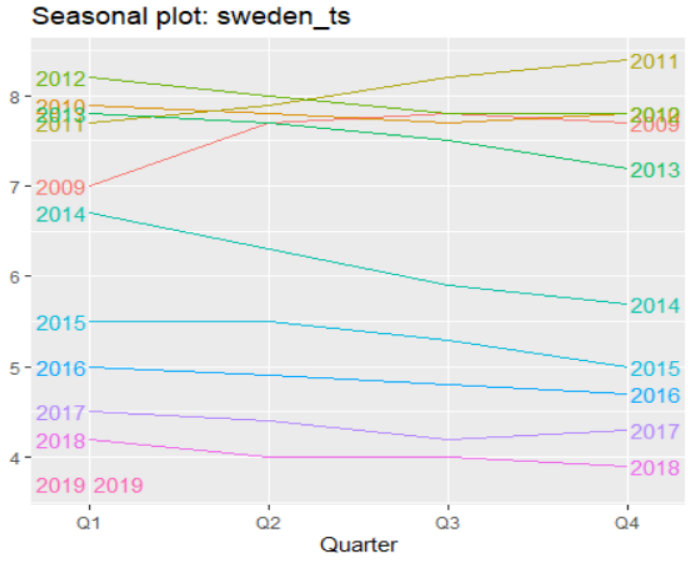
Data should be stationary to proceed with time series analysis. In order to make the data stationary, Difference function (diff) can be used. The difference value usually is one because the first difference model will be apt for the data. After plotting the differenced data, from the graph below we can see that the data is stationary with no trend. That implies that data is independent of time.

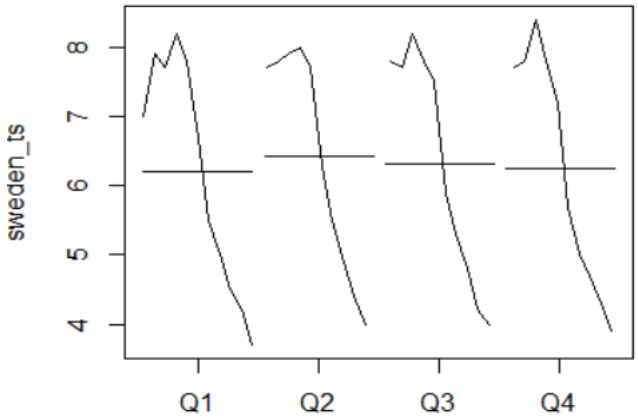
```
# Differencing data
DY=diff(sweden_ts,differences = 1)
DY
plot(DY)
```



By using seasonal plot graph, we can analyze the different fluctuations of the data in the given time periods (quarterly). From 2009 to 2019 for each quarter the data is plotted seasonally. The trend in the data is clearly seen in the below graph. 'monthplot' function also enables us to see the trend not only for monthly data also for quarterly distributed data. Apart from fluctuations, from the year 2014, there is a continuous decrease in the unemployment percentage. The graph is shown below.

```
ggseasonplot(sweden_ts,year.labels = TRUE,year.labels.left = TRUE)
monthplot(sweden_ts)
```



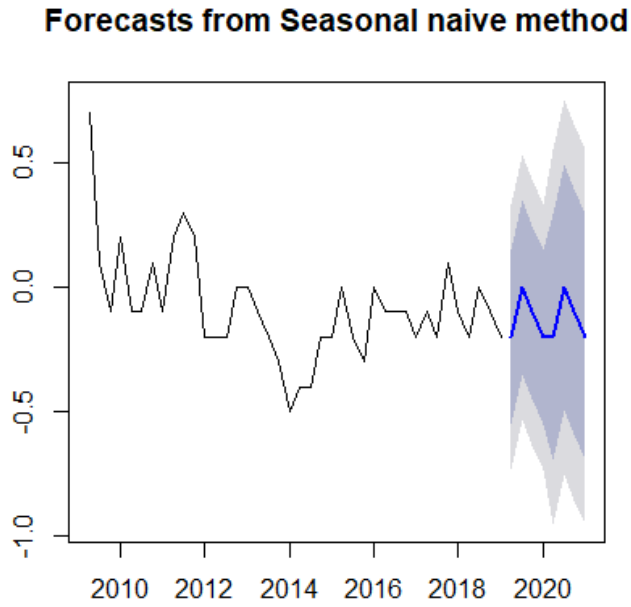


MODEL EVALUATION AND FORECASTING:

1.NAIVE MODEL:

It predicts future values based on the previous actual values and it can take both seasonality and trend. By using “snaive” function we can analyz,e Naive model. Differenced data is given for naive model and then it is forecasted. Root mean square error (RMSE) value is taken from summary which is 0.2718. The naive plot implies that there is continuous upward and downward trend.

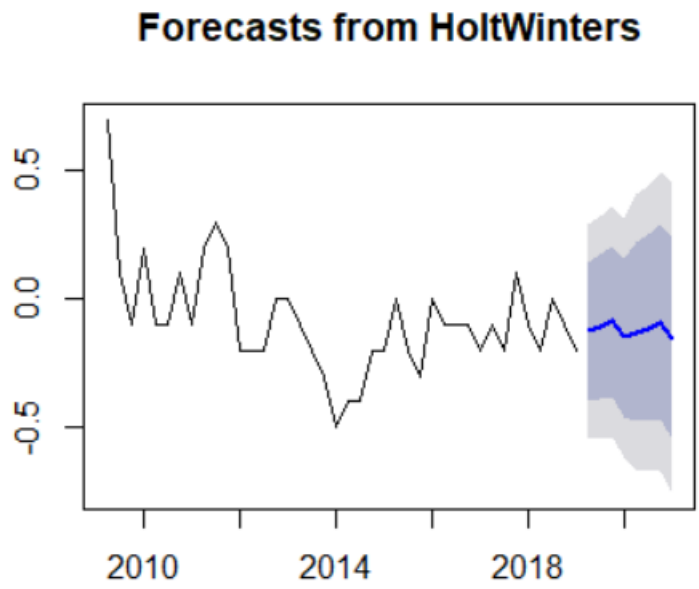
```
#Naive model
fit=snaive(DY)
print(summary(fit)) #RMSE:0.2718
checkresiduals(fit)
s=forecast(fit)
plot(s)
```



2.HOLT-WINTER MODEL:

Holt winters model is performed on the stationary data and it is plotted by using plot function. From forecast graph we can observe the changes between the confidence intervals. From summary of the forecast, there is decrease in the percentage of unemployment from 3.59 to 2.33. The RMSE value is 0.2039 which denotes it is slightly less than the previous naive model.

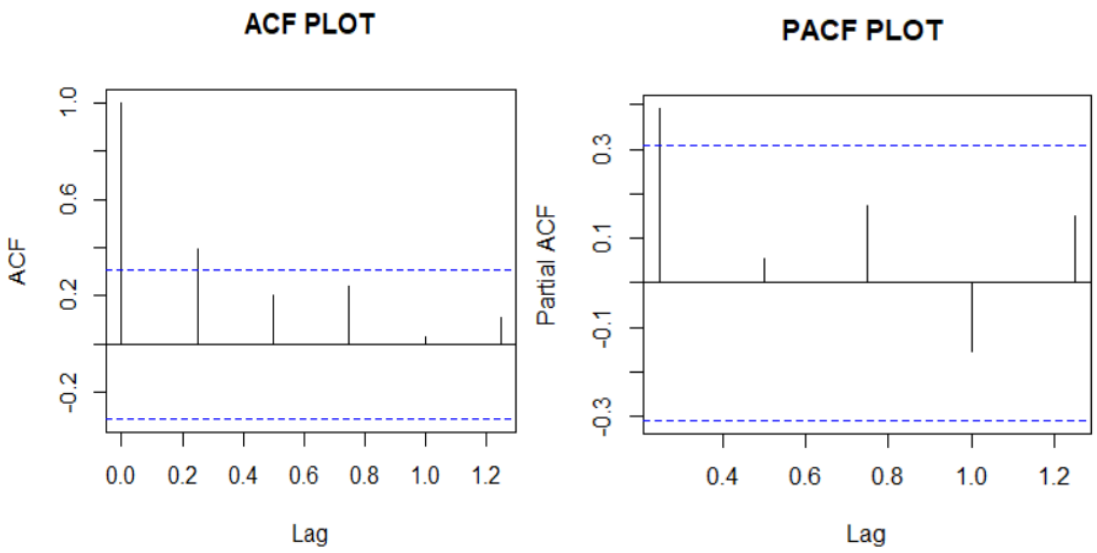
```
#Holt-winters Method
hw=HoltWinters(DY)
plot(hw)
summary(hw_forecast)# RMSE :0.2039
plot(forecast(hw))
```



3.ARIMA MODEL:

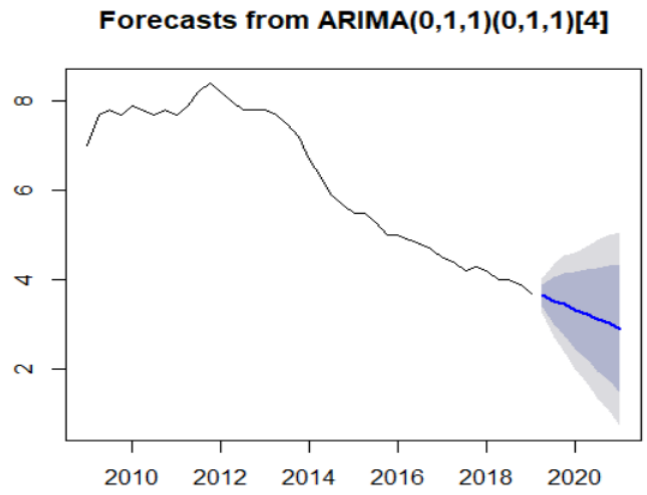
First, by manually plotting, autocorrelation (ACF) and partial autocorrelation (PACF) factors were analyzed. Then they are plotted for calculating the ‘p’ and ‘q’ values. By using ‘ACF’ and ‘PACF’ functions correlation factors are calculated with lag or delay of 5. From the Acf plot there are two spikes and from Pacf plot there is one spike in the upper part of the graphs. That implies that this is Auto regression model. So acf and pacf values are ‘p’ and ‘q’ respectively. As there is a gradual decrease in spikes of acf plot and spikes in pacf plot denotes that it belongs to AR(1) model. For calculating the correlation factors, it is then given in order. In summary, RMSE is obtained as 0.1787. From the forecast graph, straight line with slight deviations is observed.

```
# Manual ARIMA Model
ac=acf(DY, lag=10, main='ACF PLOT')
plot(ac)
pc=pacf(DY, lag=10, main='PACF PLOT ')
plot(pc)
manual_arima=arima(sweden_ts,order = c(2,1,1)) #RMSE:0.1787
summary(manual_arima)
```



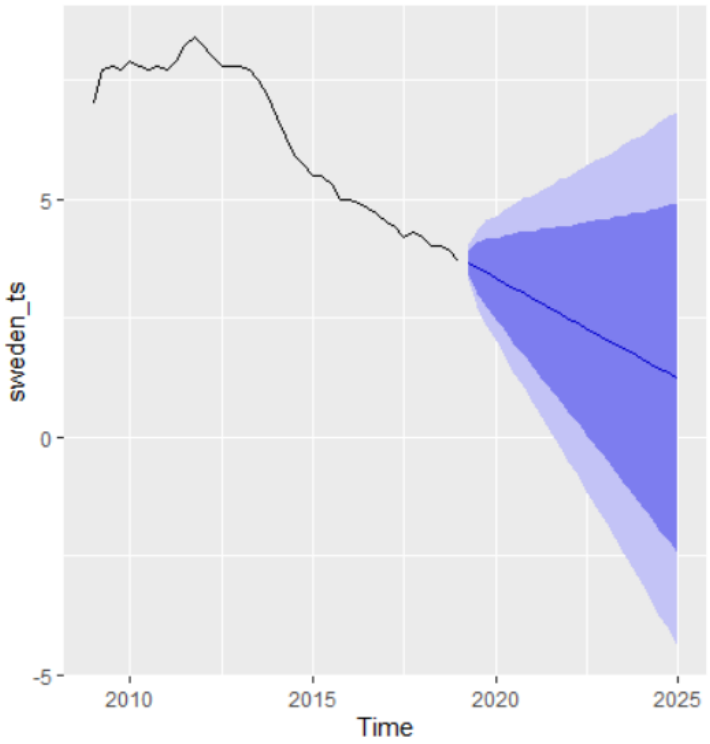
Secondly, by using 'autoarima' function plotting and analyzing. It can perform on non-stationary data. So time series data is fed to the model. This function will automatically evaluate all the possible small and capital 'p' and 'q' combinations and chooses one best model out of those. That may eventually differ from the manual calculation done above. In the function there are different 'd' values, in which small 'd' represents the differencing where the first difference is taken for this data. And the 'D' represents the seasonality difference that is taken as one. By taking the seasonal difference, we can get rid of seasonality. By making stepwise as false, thorough search will be done to find out the best model. Autoarima model chooses (0,1,1) as best model out of various combinations. The RMSE value for auto arima model is 0.1759 which is less than all previous models.

```
# Auto ARIMA Model
fit_arima=auto.arima(sweden_ts,d=1,D=1,stepwise = FALSE,approximation = FALSE,trace = TRUE)
fit_arima
print(summary(fit_arima)) # RMSE:0.1759
checkresiduals(fit_arima)
fc=forecast(fit_arima)
plot(fc)
```



FORECASTING BEST MODEL AND CONCLUSION:

By taking Root mean square error (RMSE) into consideration, the model which has low or less RMSE value is said to have high accuracy among the different models. As Autoarima model has a low RMSE value (0.1759) that is taken as the best model and it is forecasted to predict the future values accurately. From the forecast of upcoming five years graph we can see that it is continuing a linear downward trend in the percentage of unemployment in Sweden. The light and dark blue colours indicate 95 and 80 percentage of confidence intervals. As a conclusion, there is a gradual decrease in unemployment percentage in Sweden from 3.31% in 2020 Q1 to 1.34% in 2024 Q4.



Forecasts:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2019 Q2	3.655720	3.40725091	3.904189	3.2757194	4.035720
2019 Q3	3.538449	3.00842172	4.068476	2.7278424	4.349055
2019 Q4	3.458396	2.75120420	4.165588	2.3768395	4.539953
2020 Q1	3.318949	2.47092115	4.166976	2.0220024	4.615895
2020 Q2	3.232598	2.23466334	4.230533	1.7063885	4.758808
2020 Q3	3.115327	1.96237805	4.268276	1.3520437	4.878610
2020 Q4	3.035274	1.74581383	4.324735	1.0632146	5.007334
2021 Q1	2.895827	1.48311362	4.308541	0.7352682	5.056386
2021 Q2	2.809477	1.26060289	4.358350	0.4406786	5.178275
2021 Q3	2.692205	0.99803454	4.386376	0.1011946	5.283216
2021 Q4	2.612153	0.78419765	4.440108	-0.1834634	5.407769
2022 Q1	2.472706	0.52025742	4.425154	-0.5133062	5.458717
2022 Q2	2.386355	0.29700263	4.475708	-0.8090337	5.581744
2022 Q3	2.269084	0.03347457	4.504693	-1.1499855	5.688153
2022 Q4	2.189031	-0.18383709	4.561899	-1.4399577	5.818020
2023 Q1	2.049584	-0.45287137	4.552039	-1.7775912	5.876759
2023 Q2	1.963233	-0.68049009	4.606957	-2.0799928	6.006460
2023 Q3	1.845962	-0.94809251	4.640017	-2.4271757	6.119100
2023 Q4	1.765910	-1.17079084	4.702610	-2.7253861	6.257205
2024 Q1	1.626462	-1.44610912	4.699034	-3.0726302	6.325555
2024 Q2	1.540112	-1.67934479	4.759569	-3.3836222	6.463846
2024 Q3	1.422841	-1.95219175	4.797873	-3.7388260	6.584507
2024 Q4	1.342788	-2.18095800	4.866534	-4.0463164	6.731892
2025 Q1	1.203341	-2.46293004	4.869611	-4.4037366	6.810418