## **QMM 4520 / 5520**

**Joshua Hubbard**

**Vasudha Gulati**

**Rohith Kumar Diwakar**

## Our forecasting goal is to be able to accurately predict the United States GDP for the next two quarters using the last 35 years of data from the St. Louis Federal Reserve

## **Scenario 1: Multiple Regression using three independent variables:**

## Consumer Price Index (for all goods and services)

## U.S. Exports (to all countries)

## U.S. Imports from China (largest importer by US as of 2022(1))

## **Scenario 2: Find the best fit model for GDP(Y) and Date(X)**

## **Compare and pick best scenario to forecast United States GDP**

Data Source: https://fred.stlouisfed.org

* United States Gross Domestic Product (GDP) - $Mils / Qtrly / Not SA
  + U.S. Imports of Goods by China / $Mils / Monthly Converted to Qtrly / Not SA
  + U.S. Exports of Goods to World / $Mils / Monthly Converted to Qtrly / Not SA
  + U.S. Consumer Price Index for all Goods / %Growth / Monthly Converted to Qtrly / Not SA

[Link to Project Presentation](https://docs.google.com/presentation/d/1aw9KfrxS3VT8zwy5545wAzUKqpCuo8Wfr1rGQpGFKtc/edit?usp=sharing)

**Index**

[Analysis of CPI (Consumer Price Index)](#_k3sjbqdgzee)

[Analysis of U.S. Exports Variable](#_ihty13p99q97)

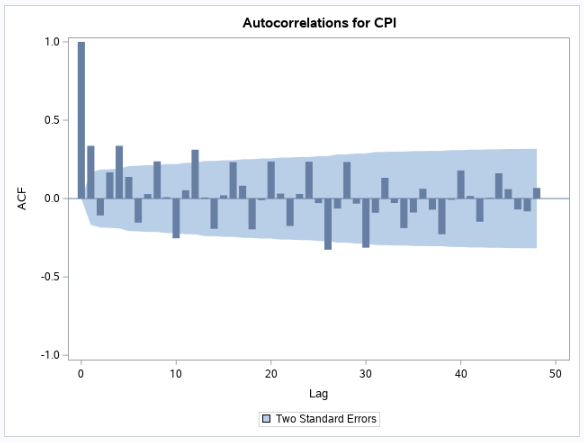
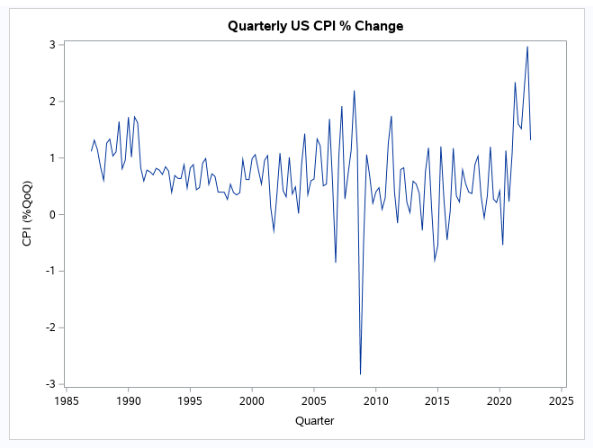
[Analysis of U.S. Imports from China Variable](#_ixycu6h4f4zl)

[Analysis of U.S. GDP](#_8rzaqu4bl40j)

[Conclusion](#_4p8w36xr8epv)

[SAS CODE](#_7tztjyi4tb7l)

## **Analysis of CPI (Consumer Price Index)**

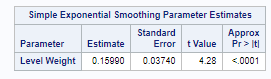
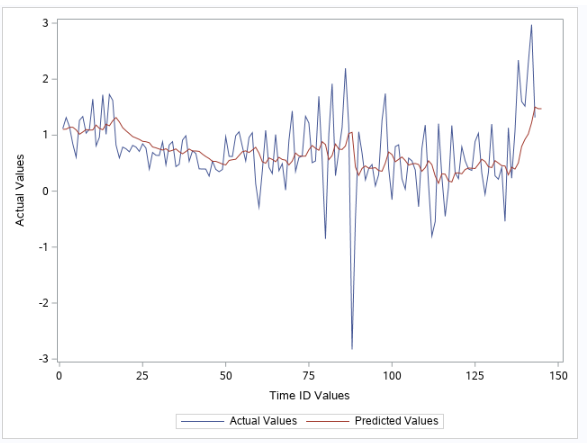


After studying a simple time series and ACF plot on the data, it appears that there are no trend or cyclical components in the data and it seems very stationary. However, there are plenty of swings / noise in the data in between. The ACF plot confirms the lack of trend, but confirms that seasonality exists.

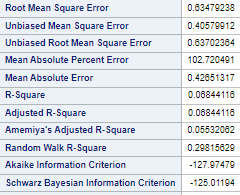
Since the model really only shows clear signs of seasonality, we have performed the following models to check to see which would be the best fit for forecasting CPI

1. Simple Exponential smoothing
2. Moving Averages
3. Winter’s Model
4. ARIMA

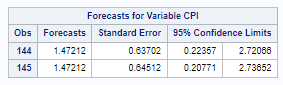
**Simple Exponential Smoothing (CPI)**



The level smoothing parameter value of 0.15 indicates that less weight is assigned to the most recent observation. The Simple Exponential smoothing model does not take into account seasonality so this is a contributor to the poor fit as displayed in the plot and explains why it does not take into account the seasonal spikes.

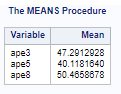


The RMSE of 0.63 and MAPE of 102 are not very good for a fit. We will use this to compare other models

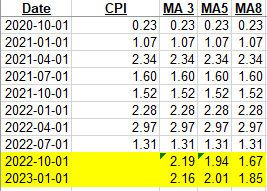


Forecast given for the next two quarters from Simple Exponential smoothing

**Moving Averages (CPI)**

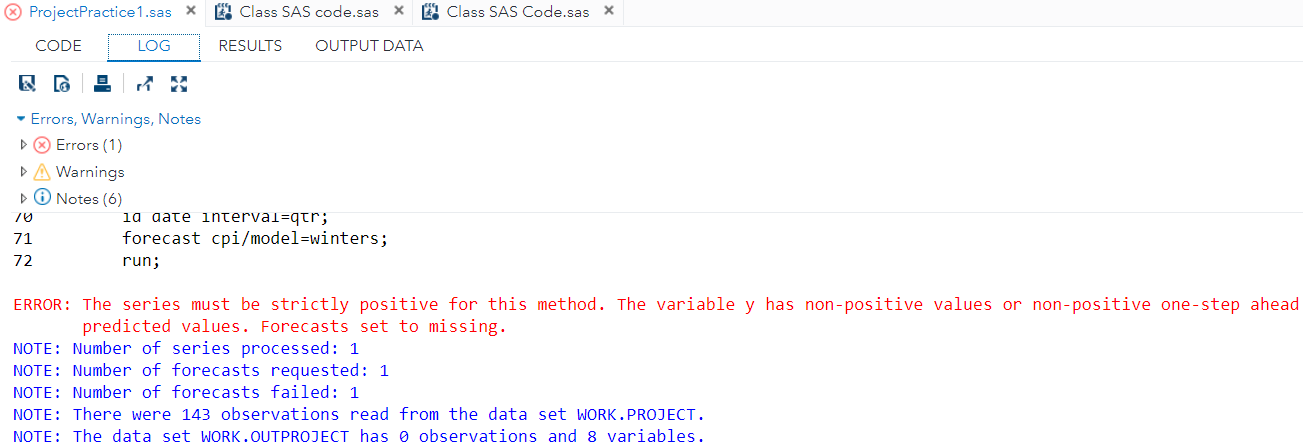


We decided to calculate three different moving averages. One is short of a full season, one includes a full season, and the third is two full seasons. After calculating 3, 5, and 8 period moving averages, the above fit is applied. The MAPE for MA5 is the best fit out of the three. So far, the MA methodology seems to capture the fit better than the simple exponential smoothing model above as it captures the spikes / seasonality better. The MAPE being much lower at 40 vs. 102 verifies this. Below are the forecasted values for the next two quarters using the three moving averages.

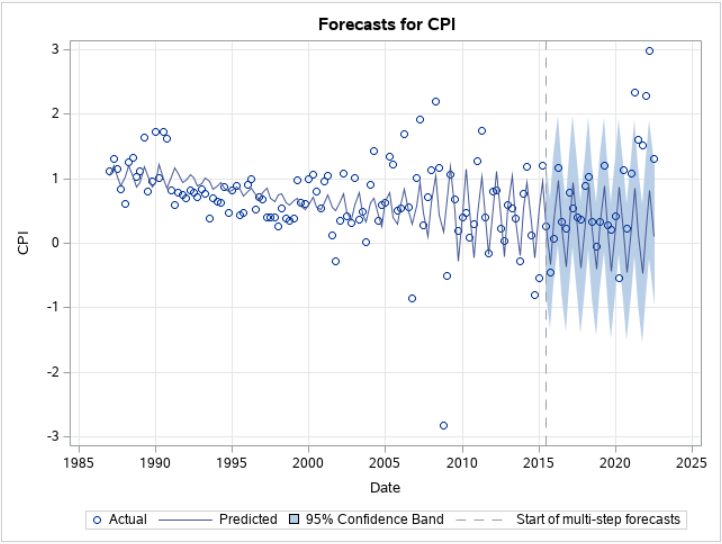


**Winters’s Exponential smoothing (CPI)**

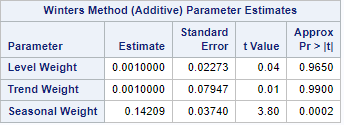
**Multiplicative Method:**

Multiplicative method did not work because CPI values contain negative values.

**Additive Method:**

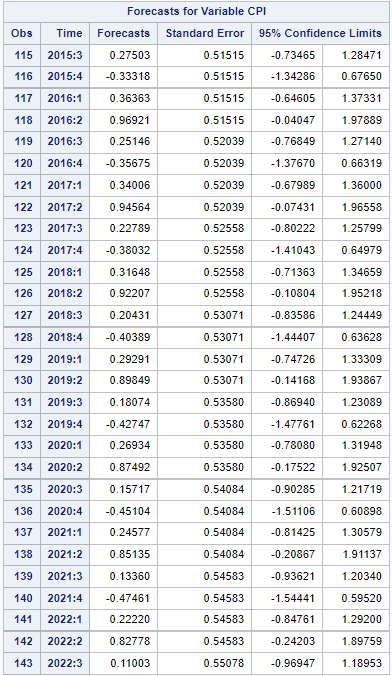
****

Winter’s exponential smoothing model has three smoothing parameters (level, trend and seasonality)

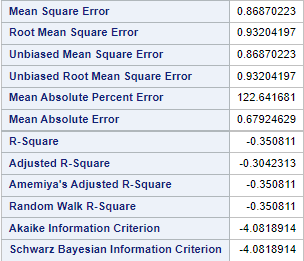
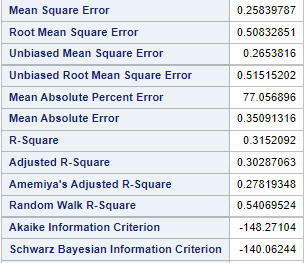


The level smoothing parameter value of 0.001 indicates that less weight is assigned to the most recent observation. The trend smoothing parameter value of 0.001 indicates that the slope is hardly changing. The seasonal smoothing parameter value of 0.14 indicates that seasonality is hardly changing.

**Forecasted Values:**

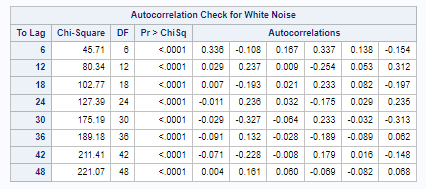


**Metrics for Model FIT vs Model Accuracy:**

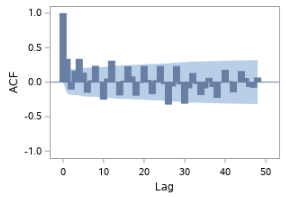
****

**Seasonal ARIMA (CPI)**

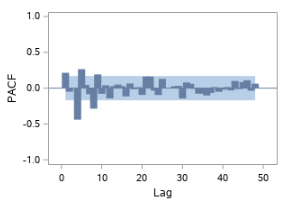
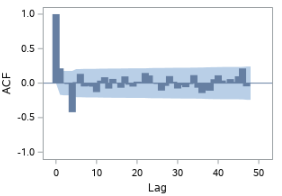
The CPI data is not white noise because the p-value is less than alpha for all the lags.

****

Below ACF and PACF plots show that the data is not stationary and ACF depicts seasonality with no trend. Therefore, we will apply seasonal differencing (4).



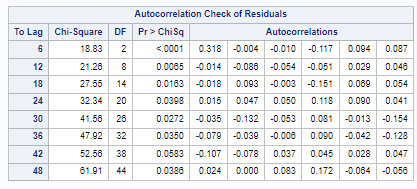
After applying differencing, the data looks stationary as there are no more seasonal trend spikes on the quarterly lags:



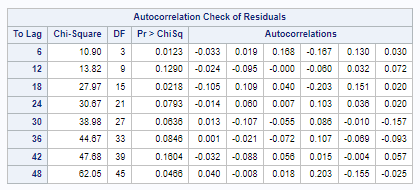
2 spikes in the PACF plot at the 4th and 8th lag so P = 2. There are no spikes in the ACF plot on the quarterly lags so Q = 0. For non seasonal, I see three non seasonal spikes on the PACF plot so p = 3… and two spikes on the ACF plot so q = 2

ARIMA (3,0,2 )(2,1,0)4

After running the first model, the residuals are all not greater than alpha which means this is not the correct model

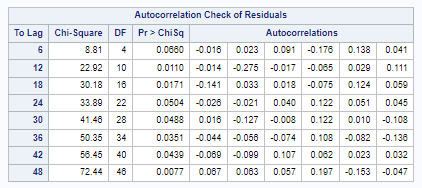


Re-analyzing the ARIMA model, I see a different variation where we only chose to have MA type model as the PACF plot seems to have more of a pattern. Therefore, we will modify our model to be ARIMA (0,0,1)(2,1,0)4



Again, not all of the residuals are whitenoise (greater than alpha). Some are better than the previous, but we must still reject the model and try another

Re-analyzing the ARIMA model, I see a different variation where we could potentially drop the second P spike in the PACF plot as it is very close to the border, therefore we will try ARIMA (0,0,1)(1,1,0)4 as our last attempt

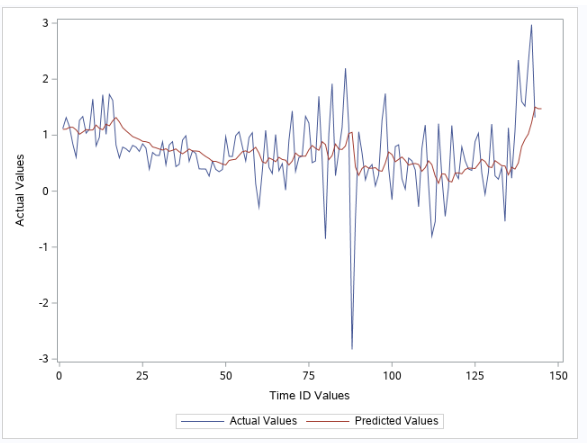


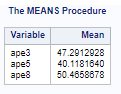
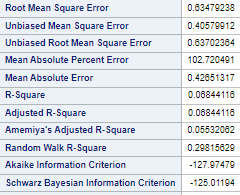
Again, not all of the residuals are whitenoise (greater than alpha). After three attempts, we will reject that ARIMA is a proper model for forecasting CPI data.

**Conclusion for CPI Variable**

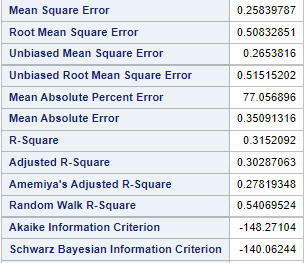
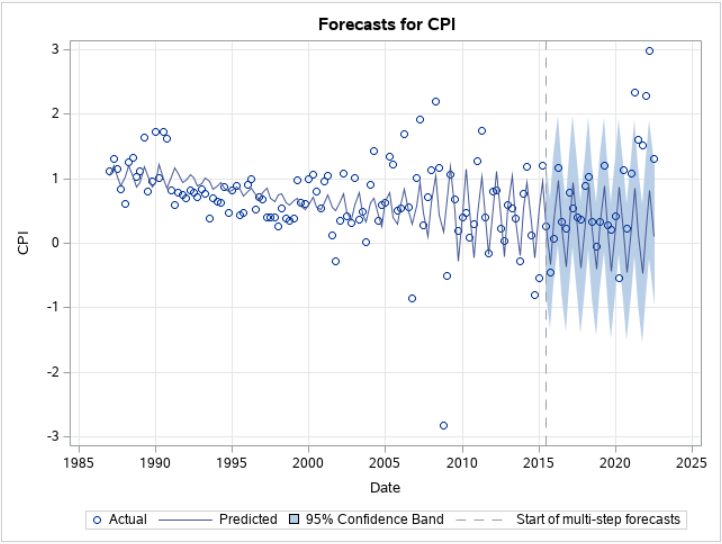
The two best models that worked with the CPI data are simple moving averages and simple exponential smoothing.

**Simple MA Simple Exp. Smoothing**

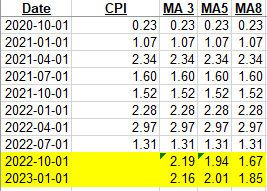


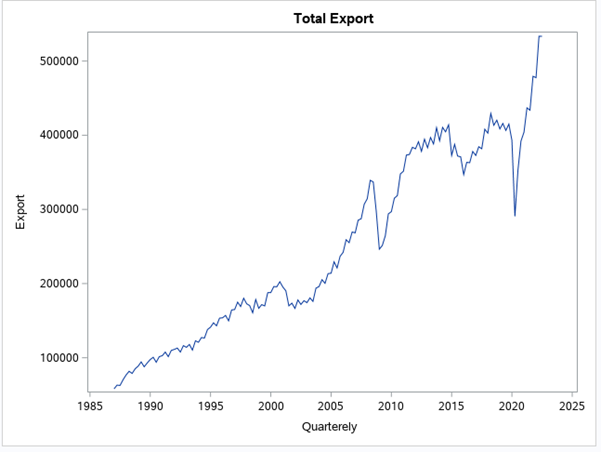
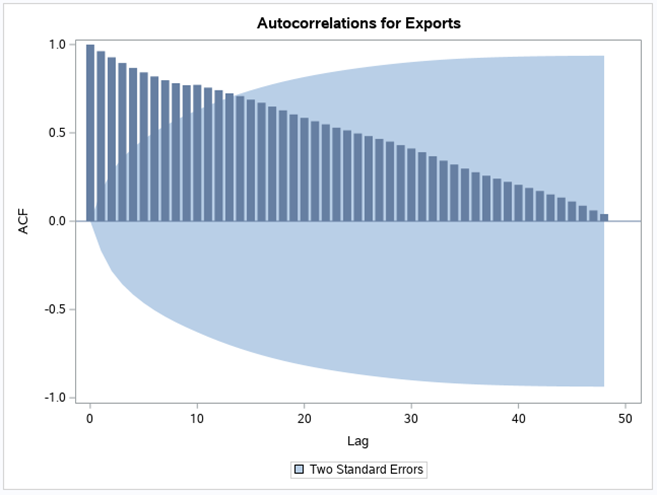
**Winter’s - Additive Model**

****

The MAPE of the 5MA is deemed as the best fit between the three models. This is confirmed visually by looking at the plots as the simple exponential smoothing model and Additive Winter’s Exponential Smoothing model do not seem to fit the spikes / seasonality components. Therefore, we will use the forecasted values output by the 5 period SMA model for our consolidated GDP model. These values are **1.94 for Q4’22 & 2.01 for Q1’23**



## **Analysis of the U.S. Exports Variable**

****

The time series plot shows a positive trend with “EXPORTS” increasing over time. This can be confirmed after considering the ACF plot.

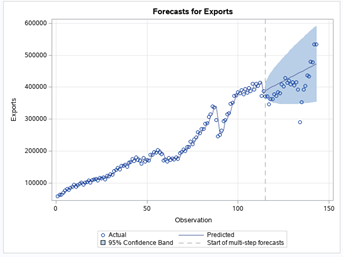
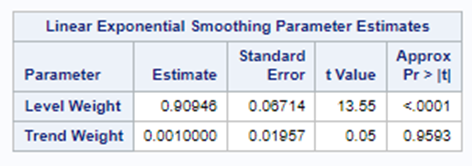
According to the ACF plot, the autocorrelations are not declining quickly toward zero, indicating the presence of a trend component.

As the data exhibit only a trend component, we initially build the model with the following:

1. Holt’s linear trend model
2. Damped trend model
3. Linear Regression
4. ARIMA

As we have 143 observations in the data, we will split the data into approx. 80:20 ratio for training and testing the data.

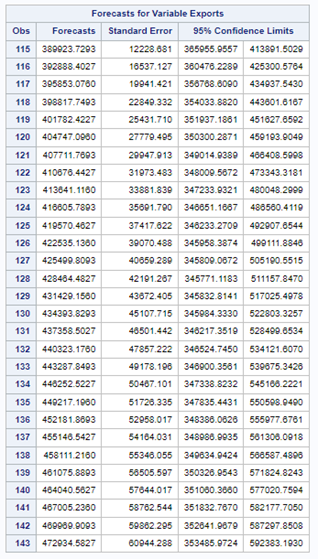
**Holt’s Exponential Smoothing**

****

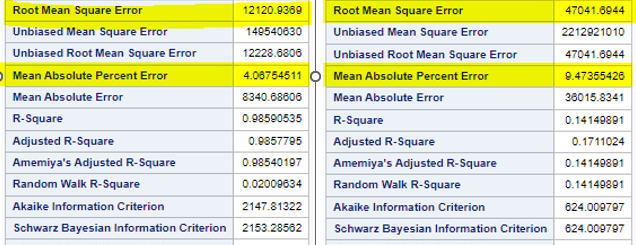
The level smoothing parameter is "**0.90946**" which indicates that higher weight is assigned to the most recent observations.

The trend smoothing parameter of "0.**0010**" indicates that the slope of the time series is hardly changing.

Below is the predicted value for test data along with the standard errors.



Below is the Statistical summary for the model fit and how well the model behaved with the new/test data.



The MAPE for model fit is **4.067%** and the MAPE for model Accuracy is **9.47%.**

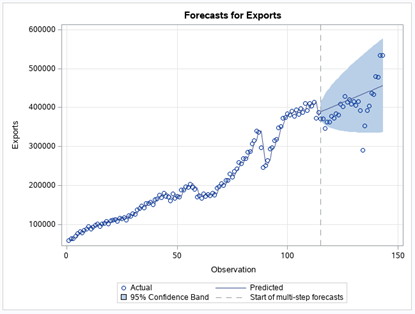
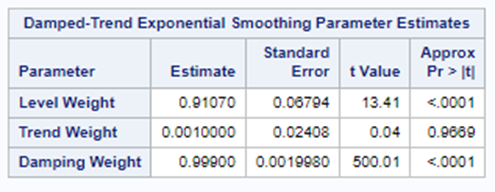
The model’s performance based on MAPE is doing a decent job, as they have low values.

The RMSE for model fit is **12121** and RMSE for model Accuracy is **47042**.

We can see a huge difference in RMSE values between model fit and accuracy. This may be due to the below reason.

We have chosen the last 29 export observations as test data. This period accounts for Exports during covid. From the time series plot, we can observe a sharp decline in Exports and a sharp rise after that. The model failed to capture these variation factors, thus explaining the higher standard error values.

**Damped Exponential Smoothing**

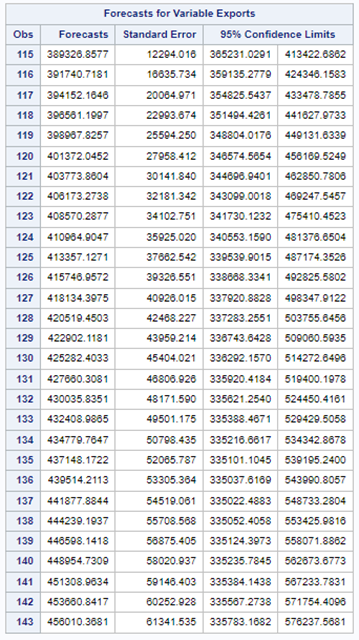
****

The level smoothing parameter is "**0.910**" which indicates that a higher weight is assigned to the most recent observations.

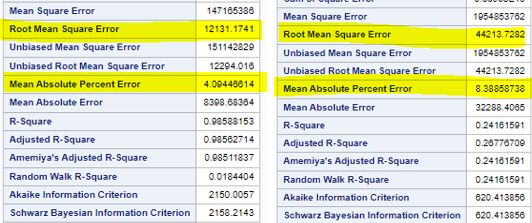
The trend smoothing parameter of "**0**.**0010**" indicates that the slope of the time series is hardly changing.

The damping smoothing parameter is "**0.99**" which indicates that NO damping is applied.

Below is the predicted value for test data along with the standard errors.



Below is the Statistical summary for the model fit and how well the model behaved with the new/test data.



The MAPE for model fit is **4.094%** and the MAPE for model Accuracy is **8.388%.**

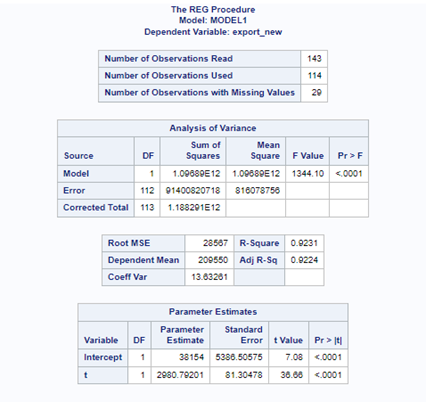
The model’s performance based on MAPE is doing a decent job, as they have low values. The model accuracy is slightly better than that of Holt’s model.

The RMSE for model fit is **1231** and RMSE for model Accuracy is **44213**.

We can see a huge difference in RMSE values between model fit and accuracy. This may be due to the below reason.

We have chosen the last 29 export observations as test data. This period accounts for Exports during covid. From the time series plot, we can observe a sharp decline in Exports and a sharp rise after that. The model failed to capture these variation factors, thus explaining the higher standard error values.

**Simple Linear Regression**

****

Equation of line fit

**Y =** 38154+ 2980.79 **\* x1**

Y = Export

x1 = t

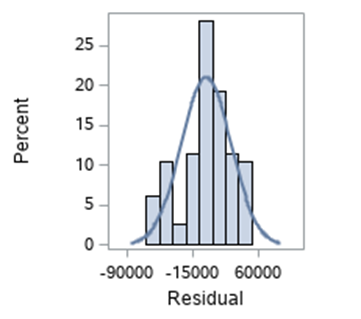
Each Quarter the Export increased by 2980.79(on average). At time = 0 Exports would be 38154 million.

Model Evaluation:

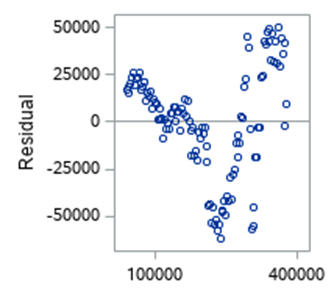
1. The model is logical because the slope is Positive which makes sense.
2. The p-value for the slope term “<0.001” is less than the alpha, so “t” is a significant predictor.
3. What percent of the variation in the dependent variable is explained by variation in the independent variable (look at R-squared). R2 = 92.31% so 92.31% of the variation in the dependent variable is explained by the independent variable

Model Assumption:

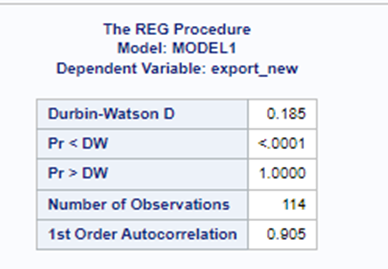
1. The relationship between the forecast variable “y” and the predictor variable “x” is linear.
2. According to the histogram, the residuals appear to be normally distributed, The assumption is true.



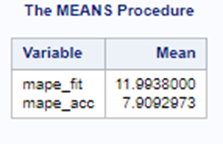
1. The shows a Fan out pattern that indicates that the equal variances assumption is not correct.



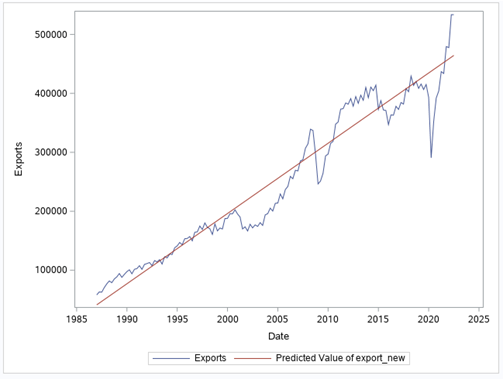
1. Independence - There is positive autocorrelation because the p-value is less than the alpha. The assumption is not true



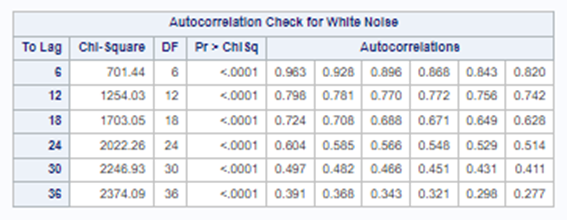
1. The model fit is 11.99% and the model accuracy is 7.909%. The Model accuracy is better than the model fit. The model accuracy of the linear regression model is better so far compared to Holts and Damped trend model.



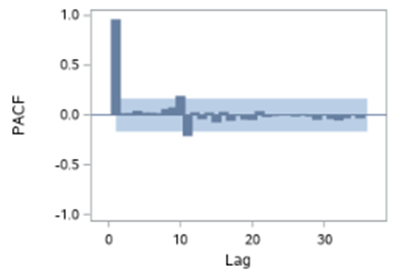
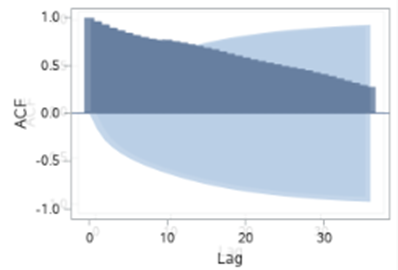
Below is the time series plot showing how well the model fit the data. We see model fit is great for the first 15 years of data.

****

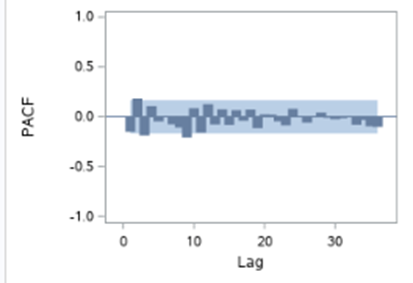
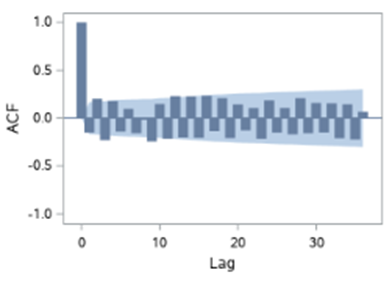
**ARIMA**

****

The data is not white noise



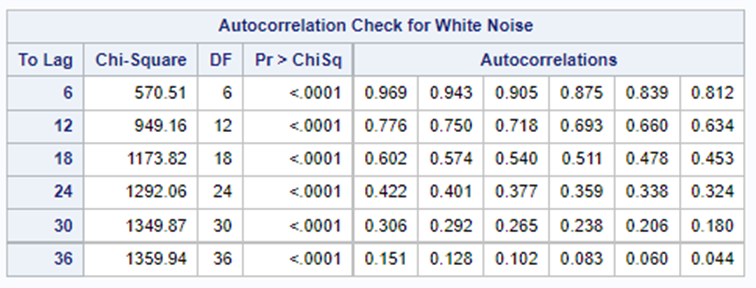
From the above charts data is not stationary, we perform first-order differencing. After first order differencing below are ACF and PACF plots.

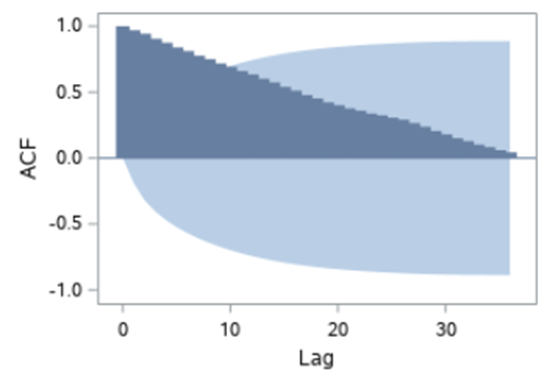


From the above ACF, plot data is not stationary. We see the sinusoidal pattern in the ACF plot and no pattern in the PACF plot. So we now identify the model.

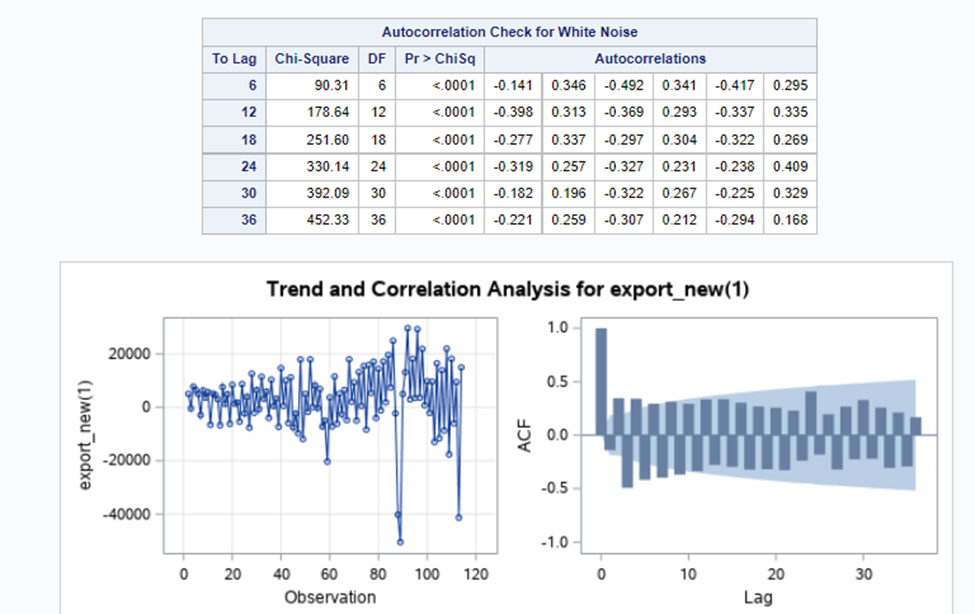
So there are two significant spikes in PACF. Our model will be ARIMA(2,1,0).

Step 1: Identify

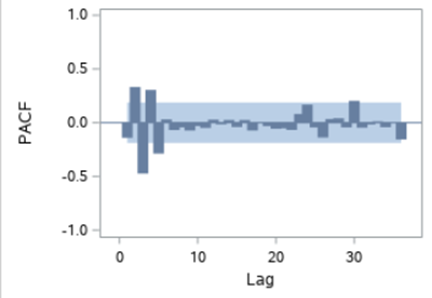
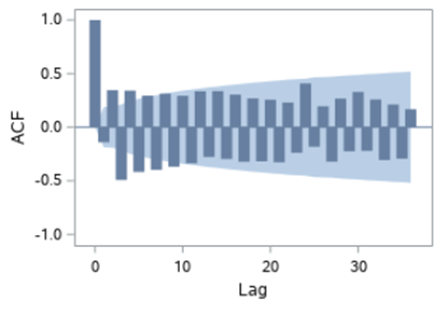




The time series is not stationary as there is no rapid decline in the autocorrelation values zero. The p-value is less than the alpha value concluding data is not white noise. So we now perform first-order differencing to convert the data into stationary.



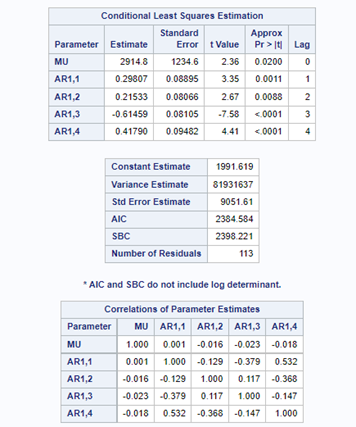
After first-order differencing, in the ACF plot, there is a rapid decline in the autocorrelation values to zero confirming the stationary.



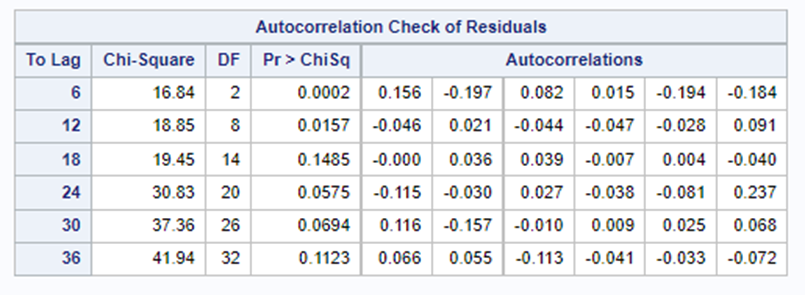
Looking at the ACF and PACF plots, there is a sine wave pattern in the ACF plot and no pattern in PACF. There are four spikes in the PACF plot beyond the confidence interval. So, we use the AR model with p = 4

Step 2: Estimate

The actual estimation of the parameters of the model is like fitting a regression model.So from the above analysis we chose model AR(4), or ARMA(4,1,0)



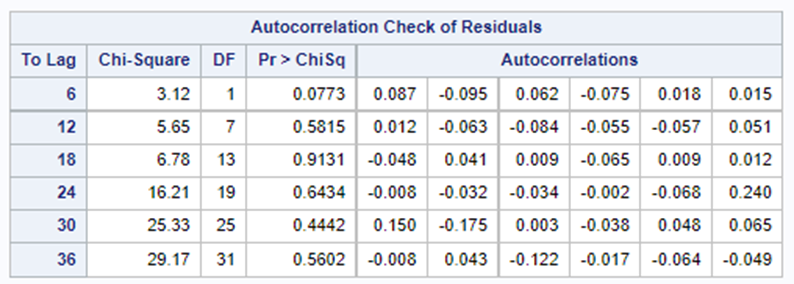
Step 3: Diagnose



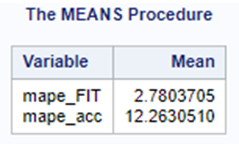
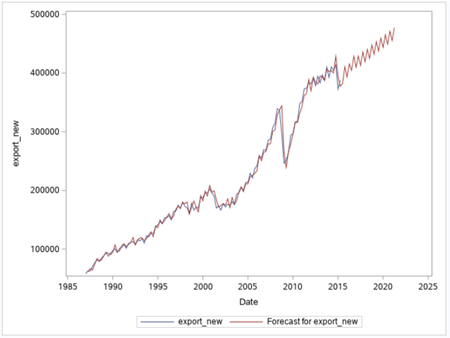
We now test whether the residual autocorrelations as a set are significantly different from zero. From the Ljung-Box-Pierce statistic, the p-values for the first two” to lags” are less than alpha, and residuals are not white noise. **Then the selected model is not appropriate.**

So we try with a higher order ARIMA model that is ARIMA (5,1,0) and check if the residual of the model is white noise.

From the Ljung-Box-Pierce statistic, the p-values for all” to lags” are greater than alpha, then the selected model is appropriate. from the below ACF and PACF plot, there are no spikes, so We can conclude that the residuals are uncorrelated (i.e. residuals are white noise)







The model fit is 2.78% and the model accuracy is 12.26%. The Model accuracy is worse than the model fit.

Our initial model without test train fit had an order “3”. This model has the order of “6”. This proves the current model is overfit.

**Conclusion U.S. Exports Variable**

| **Model** | **Model Fit** | **Model Accuracy** |
| --- | --- | --- |
| **MAPE** | |
| **Holt’s** | 4.06 | 9.47 |
| **Damped** | 4.09 | 8.38 |
| **Linear Regression** | **11.99** | **7.9** |
| **ARIMA** | **2.78** | **12.26** |

We have chosen linear regression as it has the best model accuracy of the four models ran

To determine the next two forecasted periods we can calculate using the Regression equation:

**Y =** 38154+ 2980.79 **\* x1**

38154 + 2980.79\*144 (Q4 2022) = 467387.8

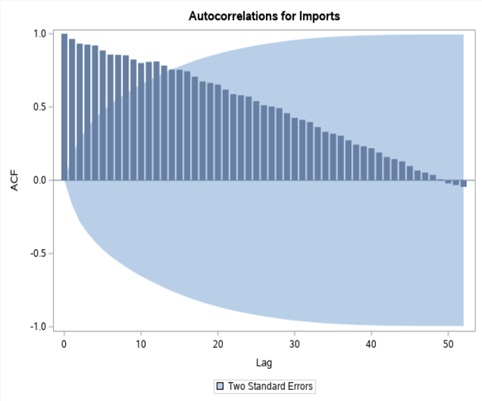
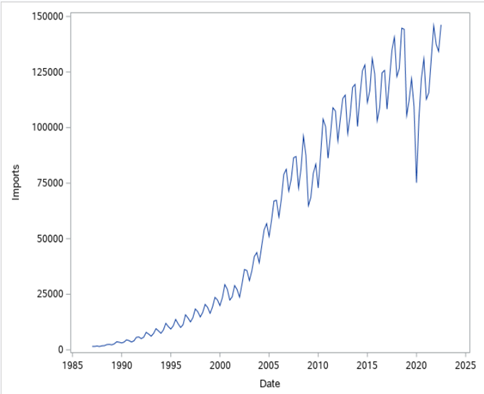
38154 + 2980.79\*145 (Q1 2023) = 470368.6

## **Analysis of U.S. Imports from China Variable**

**Time Series and ACF Plot:**

The time series plot shows a positive trend with imports increasing over the time period. We can also identify a seasonal component where imports are higher during the first and second quarters.

According to the ACF plot, the autocorrelations are not declining quickly towards zero which indicates a trend component. In addition, the autocorrelations are higher at lags 4, 8 and 12 which indicates the presence of a seasonal component.

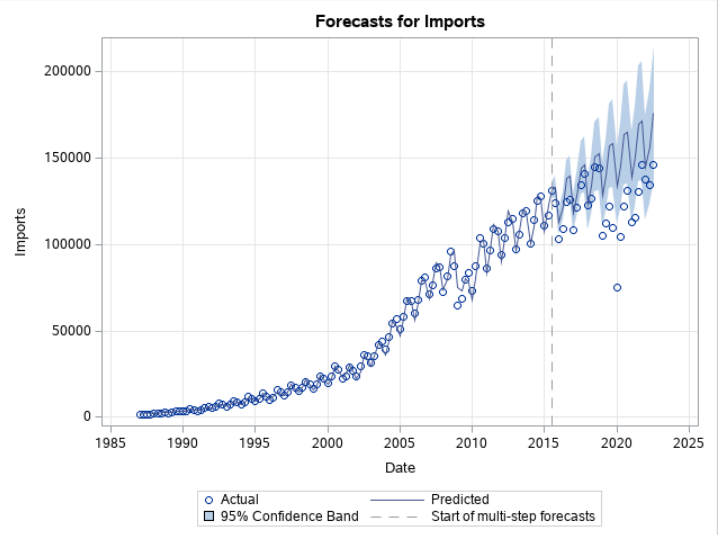


Since the data shows presence of both Trend and Seasonal Component, below methods would be appropriate:

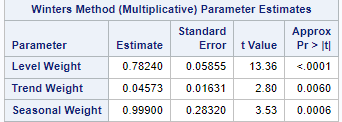
1. Winters Exponential Smoothing
2. Simple Regression Model
3. Seasonal ARIMA Model

**Winters Exponential Smoothing Method**

**Multiplicative Method:**

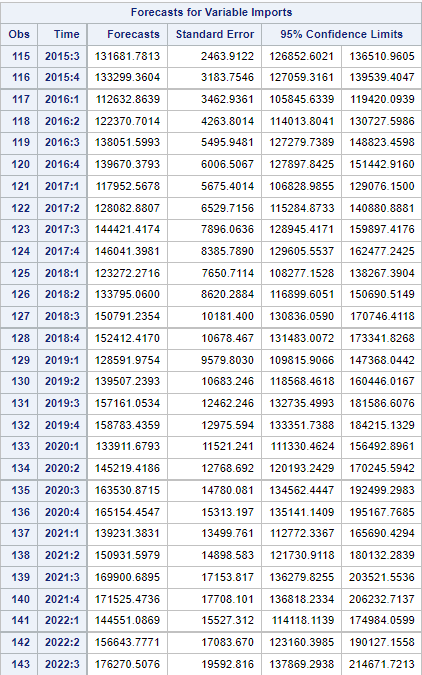
****

Winter’s exponential smoothing model has three smoothing parameters (level, trend and seasonality)

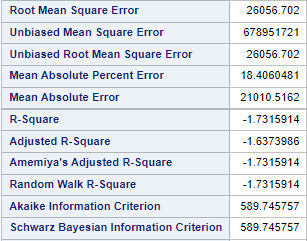
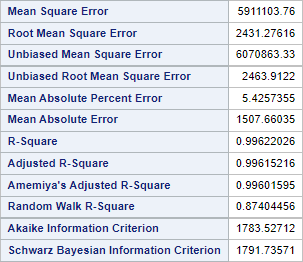


The level smoothing parameter value of 0.78 indicates that more weight is assigned to the most recent observation. The trend smoothing parameter value of 0.04 indicates that the slope is hardly changing. The seasonal smoothing parameter value of 0.99 indicates that seasonality is drastically changing.

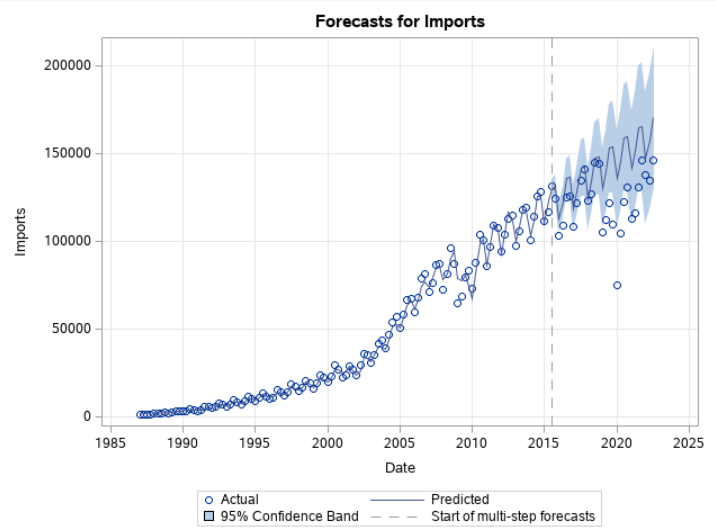
**Forecasting for next 29 quarters:**

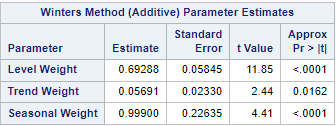
****

**Metrics for Model FIT vs Model Accuracy:**

****

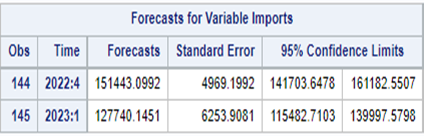
**Additive Method:**

****

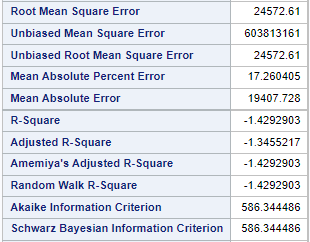
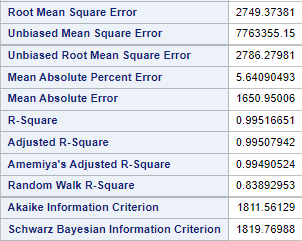
****

The level smoothing parameter value of 0.69 indicates that more weight is assigned to the most recent observation. The trend smoothing parameter value of 0.05 indicates that the slope is hardly changing. The seasonal smoothing parameter value of 0.99 indicates that seasonality is changing drastically.

**Forecasting for next two quarters and Metrics:**

****

**Metrics for Model FIT vs Model Accuracy:**

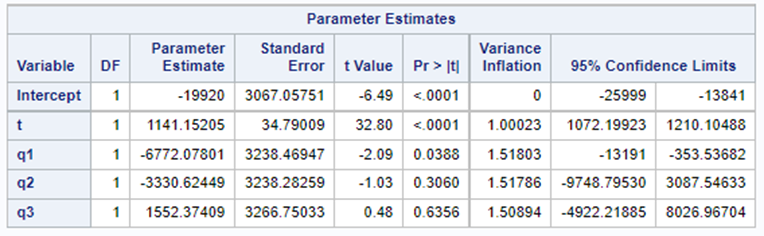
****

**Comparing both the Winter’s Models:**

| **Model: Winters** | **Multiplicative** | **Additive** |
| --- | --- | --- |
| **MAPE** | 5.42 | 5.64 |
| **RMSE** | 2431 | 2749 |
| **R-Square** | 0.996 | 0.995 |

Since the seasonal component is changing over time and also, we have lower MAPE and RMSE values for the Multiplicative method. So the **Multiplicative Method** would be a better fit.

**Simple Linear Regression Model**

****

**Equation of Line Fit**:

Y= intercept + t X1 + q2 X2 + q3 X3 + q4 X4

where y=imports, X1=t, X2=q1, X3=q2 and X4=q3

Here, Y = -19920 + 1141 X1 - 6772 X2 – 3330 X3 + 1552 X4

**Interpretation of slope parameters:**

Compared to Q4: Q1 Imports is $ 6772 million lower than Q4

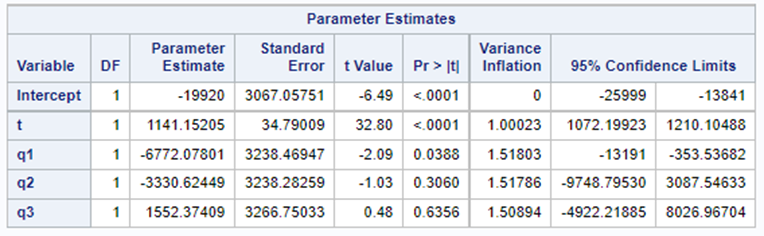
Q2 Imports is $ 3330 million lower than Q4

Q4 Imports is $ 1552 million higher than Q4

Intercept here is - Imports in Q4 is $46.12 million

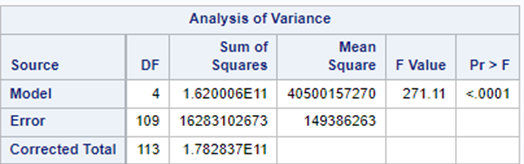
**Evaluating the Model:**

1. Model logical: Slopes of the model include positive values for t and q3 and negative for q1 and q2 which seems logical as plot depicts the same.
2. Slope terms statistically significant:



Clearly, p-value is not less than alpha for all the variables. p-value is less than alpha for t which indicates t is a significant predictor for imports and zero is not a possible value within the limits because it lies between 1072 and 1210 value. For q1, p-value is less than alpha and zero is a possible value because it lies between -13191 and -353 value. For q2 and q3 p-value is > alpha So, we can conclude, q2 and q3 are not statistically significant but t, q1 are statistically significant.

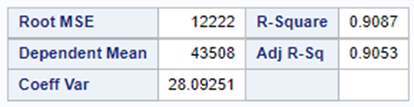
1. Model statistically significant:



P-value here shows <0.001 which shows p-value for the model is less than alpha.

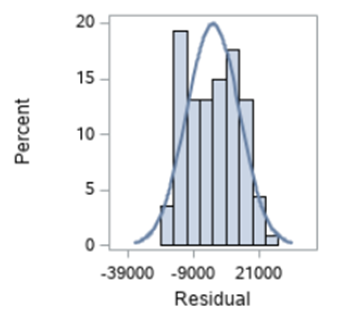
So, the model is statistically significant.

**Metrics:**



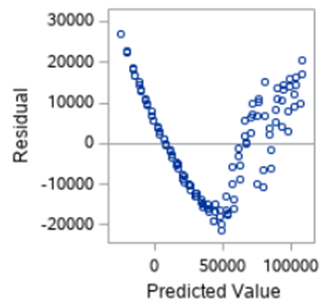
**Model Assumptions:**

1. The relationship between the forecast variable and the predictor variable is linear.
2. Normally distributed: True



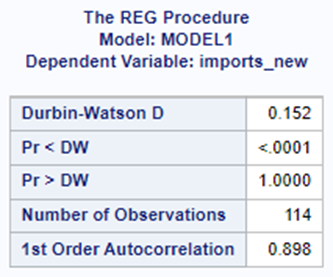
According to histogram, the residuals appear to be normally distributed because it is bell-shaped. So, this assumption is true.

1. Equal variance (homoscedastic): Not True



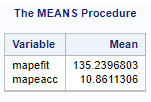
According to the scatterplot of residuals vs. predicted value, there is a pattern which so equal variances assumption is not true.

1. Independent (non-autocorrelated): Not true



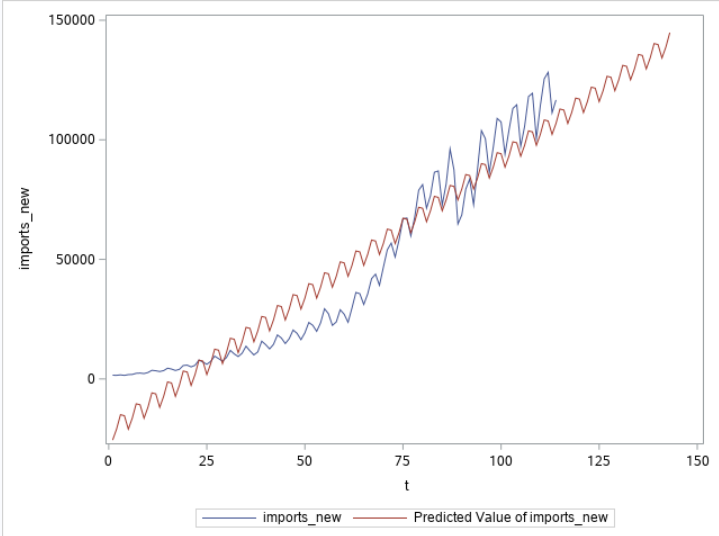
P-value for p<DW is less than alpha and p>DW is greater than alpha which shows there is positive correlation. So, this assumption is not true.

**MAPE:**



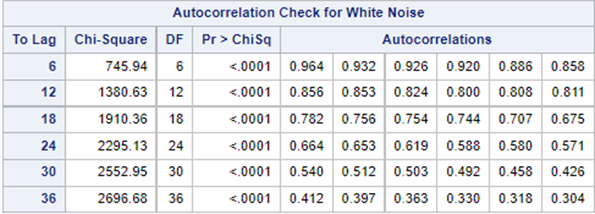
MAPE Fit is comparatively very high than MAPE Accuracy, which is kind of predictable after going through the timeseries plot, this model is not fitting the data perfectly.

**Plot:**

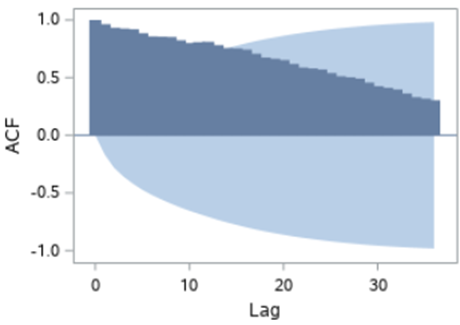
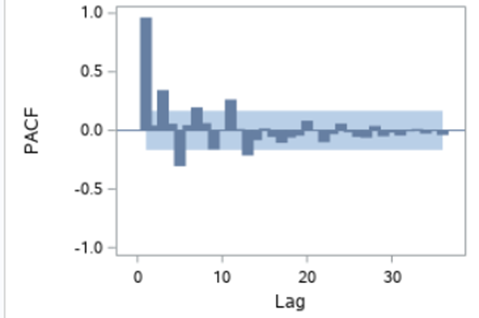


**Seasonal ARIMA**

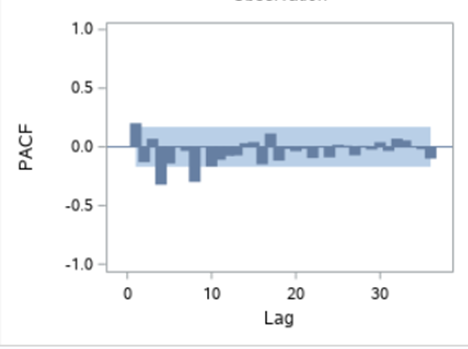
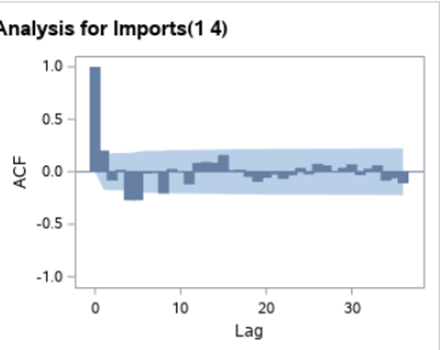
Data is not white noise because p-value is less than alpha for all the lags.



Considering the ACF and PACF plots, data doesn’t look stationary. ACF plot depicts trend and seasonality. So we will be applying (1,4) differencing.



After applying differencing, data looks stationary:



For the 1st Model, we are considering 2 spikes in PACF plot at 4th and 8th lag so P=2 and only 1 spike at 4th lag in ACF plot so Q=1. For the non-seasonal ARIMA model, we noticed 3 spikes at PACF and 3 at ACF so p=3 and q=3.

For the 2nd Model, 2 spikes in PACF plot at 4th ,8th lag so P=2 and only 1 spike at 4th lag in ACF plot so Q=1. For the non-seasonal ARIMA model, we noticed 2 spikes at PACF and 3 at ACF so p=2 and q=3.

For the 3rd Model, 3 spikes in PACF plot at 4th ,8th and 12th lag so P=3 and only 1 spike at 4th lag in ACF plot so Q=1. For the non-seasonal ARIMA model, we noticed 4 spikes at PACF and 3 at ACF so p=4 and q=3.

Below are the 3 identified models:

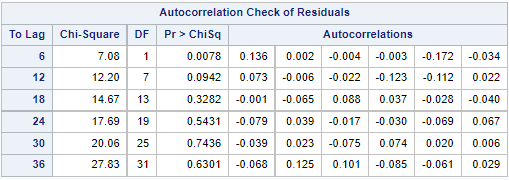
ARIMA (3,1,3) (2,1,1)

ARIMA (2,1,3) (2,1,1)

ARIMA (4,1,3) (2,1,1)

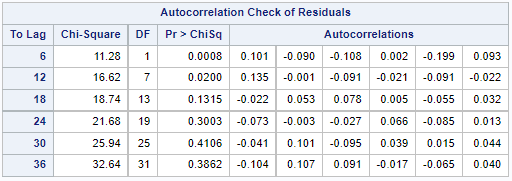
**ARIMA Model 1**: ARIMA (3,1,3) (2,1,1)

After running the first model, all the residuals except for the 6th lag are greater than alpha which means residuals are white noise and this is an acceptable model



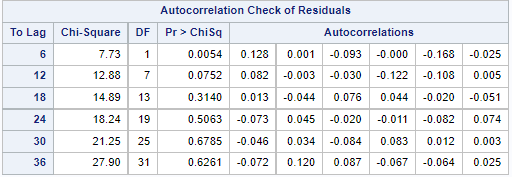
**ARIMA Model 2:** ARIMA (2,1,3) (2,1,1)

After running the second model, all the residuals are not greater than alpha which means residuals are not white noise and this is not the correct model

****

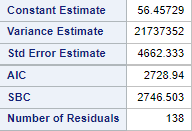
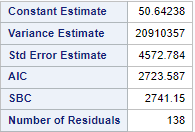
**ARIMA Model 3:** ARIMA (4,1,3) (2,1,1)

After running the first model, the residuals are greater than alpha which means residuals are white noise and this is an acceptable model



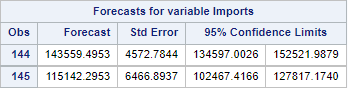
Below are the comparison between AIC and BIC values for all three ARIMA models:

**Model 1 Model 3**

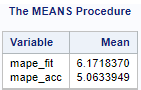


On comparing the two remaining models, they are quite similar in terms of AIC and BIC. Therefore, we will pick the model with the lowest order (Model 1)

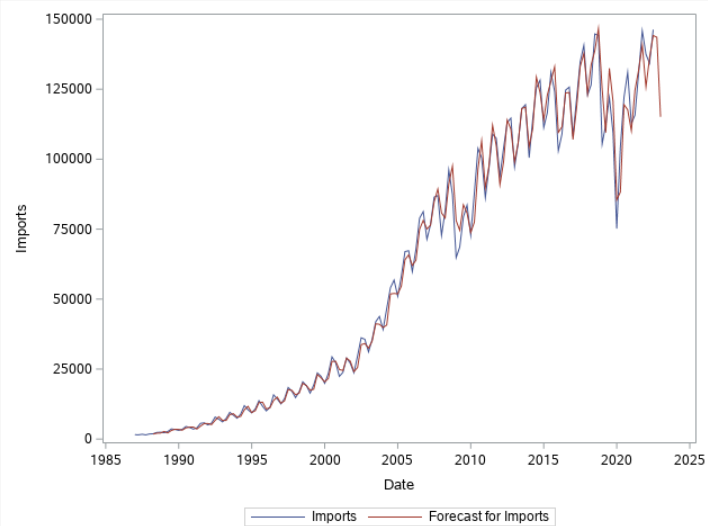
Below are the Forecasted Imports for next two quarters, if we consider Model 1:



**MAPE for Model1**:



**Plot:**



**Conclusion:**

Comparing the MAPE for all of the three models we have tried:

| **Model** | **Model Fit** | **Model Accuracy** |
| --- | --- | --- |
| **MAPE** | |
| **Multiplicative Winters** | 5.42 | 18.40 |
| **Linear Regression** | 135.2 | 10.86 |
| **ARIMA** | **6.17** | **5.06** |

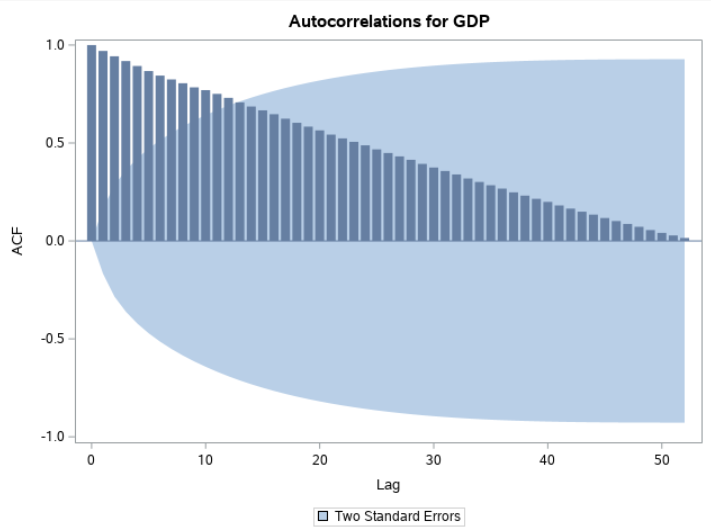
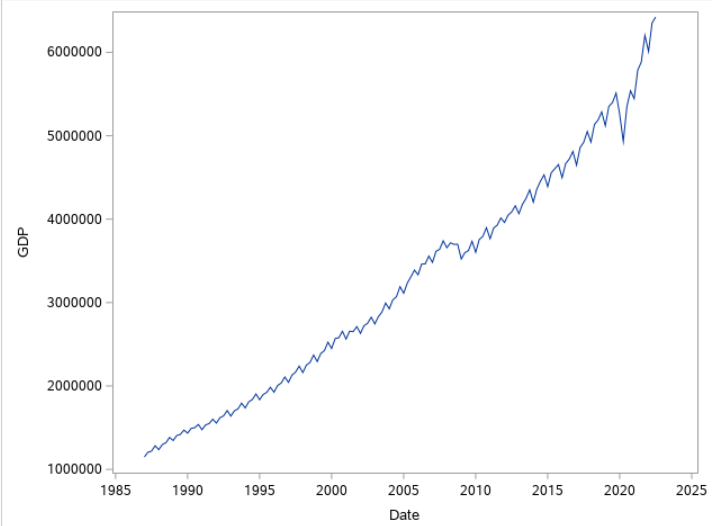
**Clearly, ARIMA is a better fit model for Imports**

## 

## **Analysis of U.S. GDP**

**Timeseries and ACF Plot:**

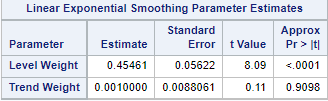
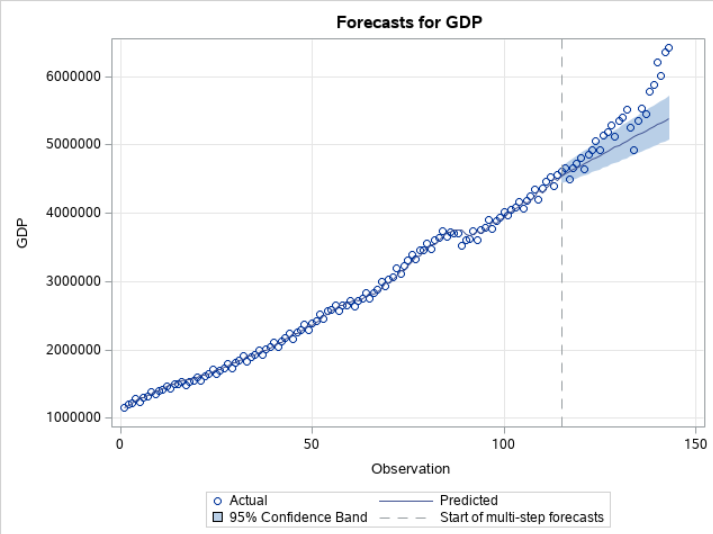
Below timeseries plot shows a positive trend and the ACF plot clearly shows autocorrelations are declining towards zero slowly which show a trend component.



Because the data exhibit only trend component, below models would be appropriate:

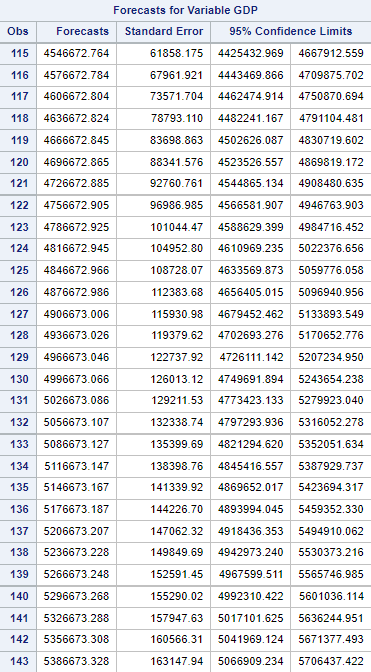
1. Holt’s linear trend model
2. Damped trend model
3. Simple Linear Regression
4. ARIMA

**Holt’s Linear Trend Model:**

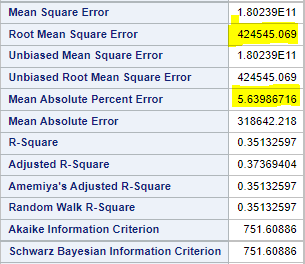
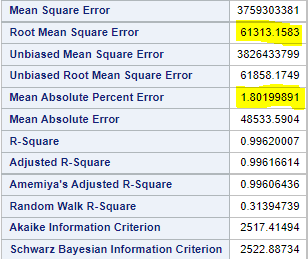


The level smoothing parameter is **0.45** which indicates that lower weight is assigned to the most recent observations. The trend smoothing parameter of **0.001** indicates that the slope of the time series is hardly changing.

Below are the forecasted value for test data:

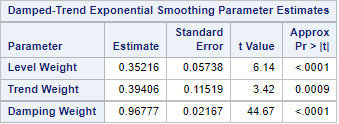
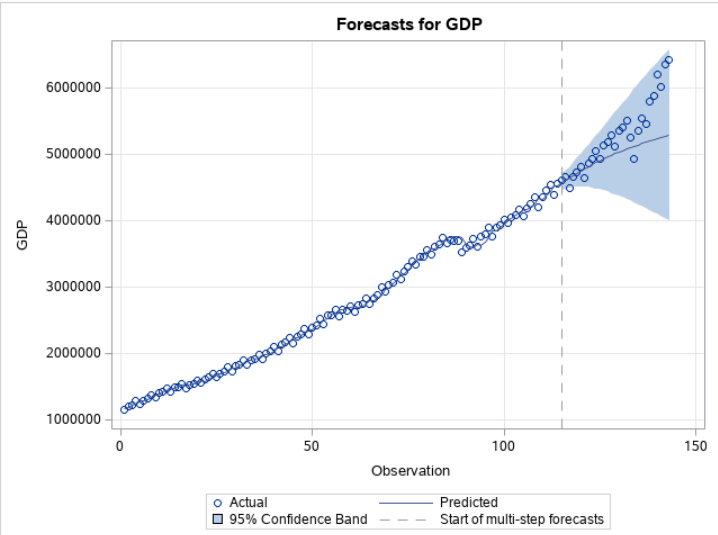


Statistical Analysis for train set and test set:



MAPE for model fit is 1.80 and MAPE for model accuracy is 5.63 which shows its a good fit model.

**Damped Trend Model:**

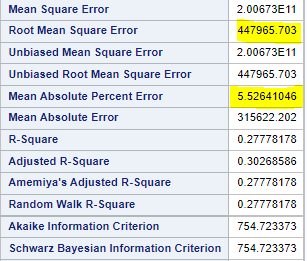
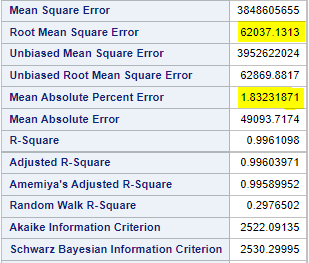
****

The level smoothing parameter is **0.35** which indicates that a lower weight is assigned to the most recent observations. The trend smoothing parameter of **0**.**39** indicates that the slope of the time series is hardly changing. The damping smoothing parameter is **0.96** which is closer to 1 and it indicates that there is hardly any damping involved.

Below is the forecasted value for test data:



Statistical Analysis for train set and test set:

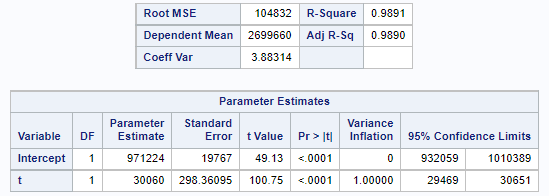


MAPE for model fit is 1.83 and MAPE for model accuracy is 5.52 which indicates its a good fit model.

| **Model** | **Holt’s** | **Damped** |
| --- | --- | --- |
| **MAPE** | **1.80** | 1.83 |
| **RMSE** | **61313** | 62037 |
| **R-Square** | **0.996** | 0.996 |

On comparing both Holt’s and Damped, **Holt’s** is a better fit model because it has lower MAPE and RMSE value.

**Simple Linear Regression**

****

**Equation of Line Fit:**

y = 971224 + 30060 x

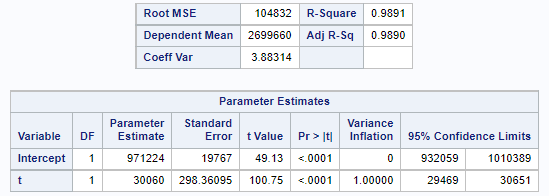
where y = gdp, x = t

Each Quarter the gdp increased by 30060(on average)

At time = 0 Exports would be 971224 million.

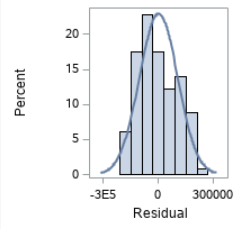
**Model Evaluation**

1. The model is logical because the slope is Positive which makes sense.
2. The p-value for the slope term “<0.001” is less than the alpha, so “t” is a significant predictor.
3. R2 = 99% so 99% of the variation in the dependent variable is explained by the independent variable

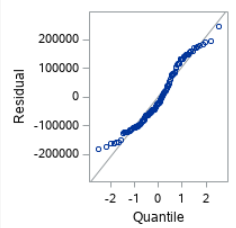
****

**Model Assumptions**

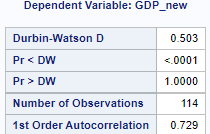
1. The relationship between the forecast variable and the predictor variable is linear.
2. According to histogram, the residuals appear to be normally distributed because it is bell-shaped. So, this assumption is true.

****

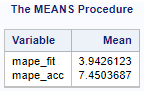
1. According to the scatterplot of residuals vs. predicted value, there is a pattern which so equal variances assumption is not true.

****

1. P-value for p<DW is less than alpha and p>DW is greater than alpha which shows there is positive correlation. So, this assumption is not true.

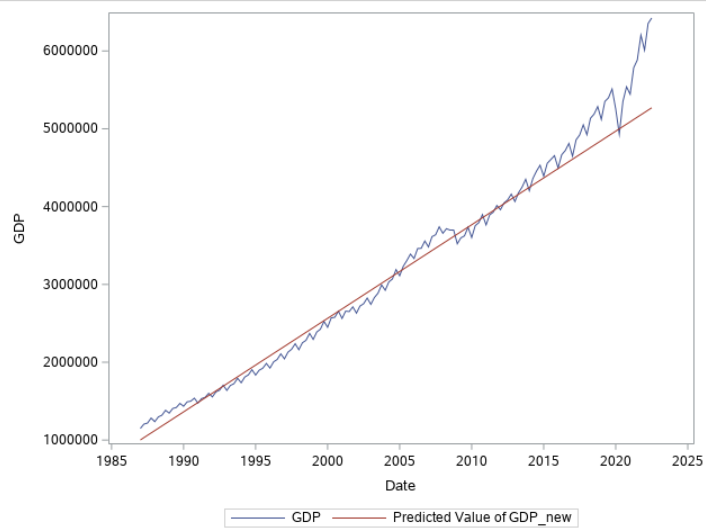
****

**MAPE:**

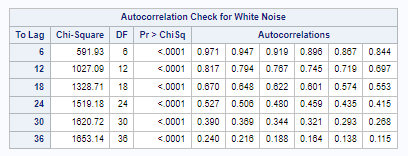
****

MAPE for model fit is 3.94 and model accuracy is 7.45 which shows its a really good model.

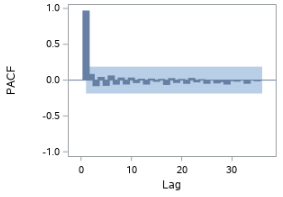
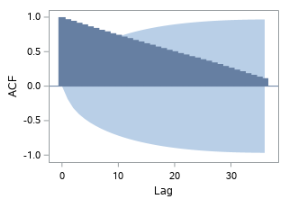
**Plot for Actual and Forecasted values:**

****

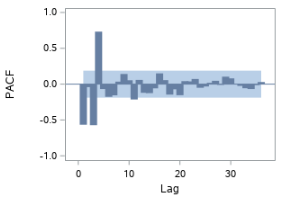
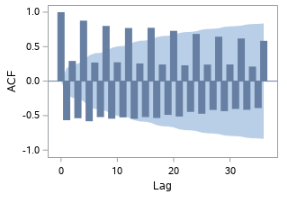
**ARIMA**

****

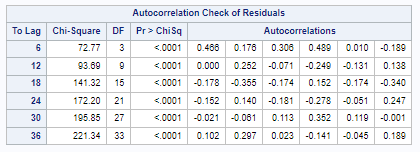
The GDP data is not white noise because the p-value is less than alpha for all the lags.



The ACF / PACF plots show that differencing is required since data is not yet stationary. No seasonality is present so we will apply first order differencing (1)



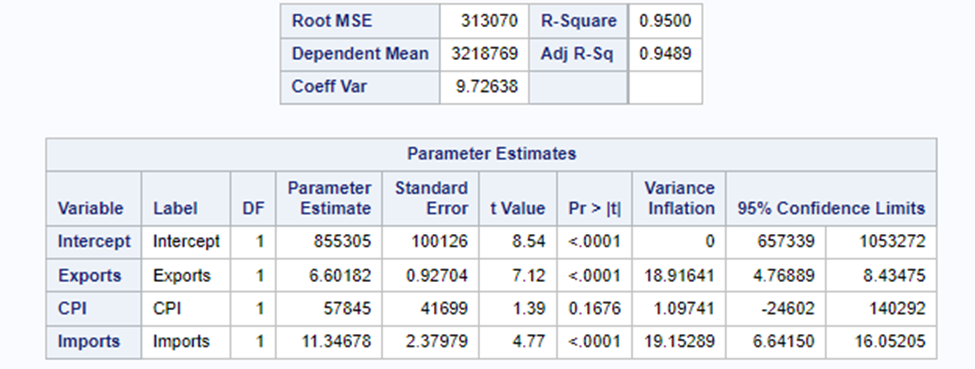
After applying first order differencing (1) - the data is now stationary. The ACF plot shows a pattern so we are looking at an AR model. There are three spikes in the PACF plot so p = 3. Arima model is then ARIMA(3,1,0)



After running ARIMA(3,1,0) our residuals did not pass the white noise test. ARIMA(2,1,0) was then run as well and we had the same result. ARIMA should not be used for GDP forecasting

**Multiple Regression**

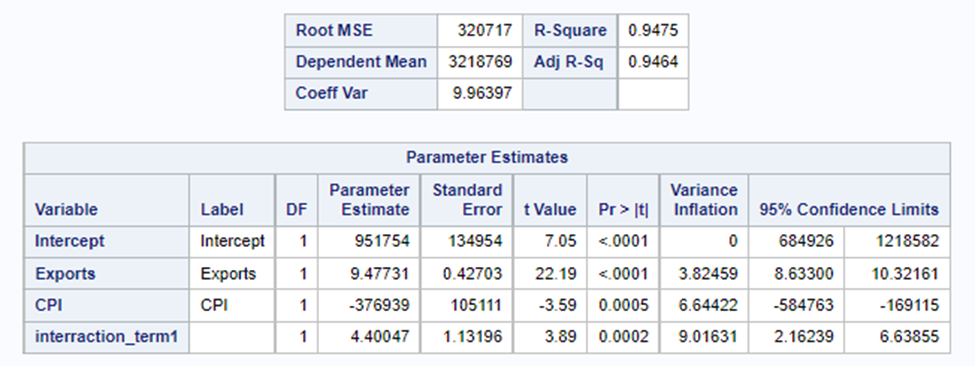
**Model 1:** - Exports, Imports, and CPI as independent variables.

****

The p-value for the term CPI is greater than the alpha value making it insignificant.

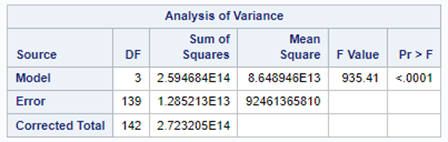
The VIF values for Exports and Imports are greater than 10 indicating the presence of multicollinearity.

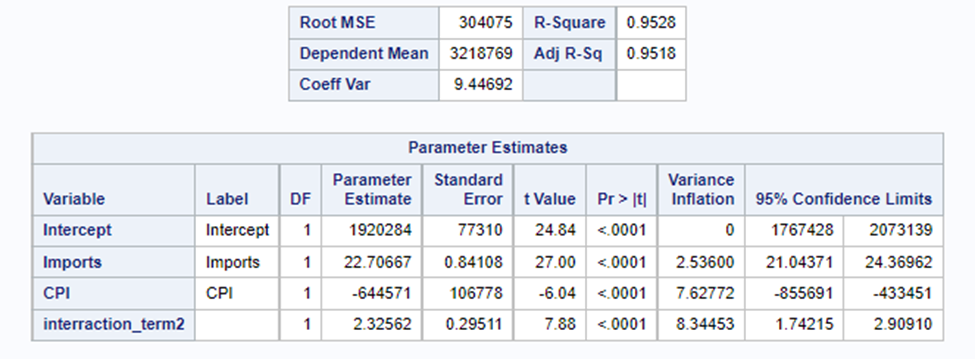
**Model 2:** - Exports, CPI and CPI\*Imports(interaction\_term1) as independent variables

****

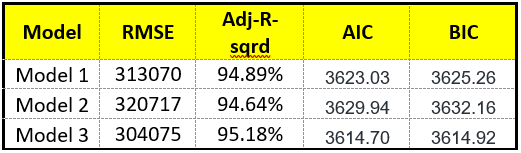
The p-values of all the variables are less than alpha making them significant. The VIF of the variables is less than 10, so there is no multicollinearity.

**Model 3:** - Imports, CPI and CPI\*Exports(interaction\_term2) as independent variables

****

****

The p-values of all the variables are less than alpha making them significant. The VIF of the variables is less than 10, so there is no multicollinearity.

****

## Therefore, model 3 has better performance.

Y= 1920284 + 22.70667\*x1 – 644571\*x2 + 2.325628\*x3

Y = GDP

x1 = Imports

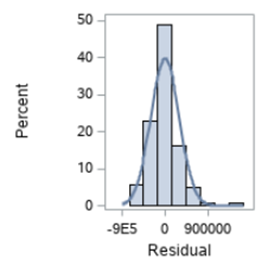
x2 = CPI

x3 = Exports\*CPI

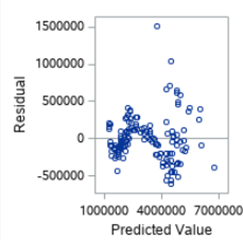
**Model Evaluation:**

1. The model is logical as the sign of all the slopes makes sense.
2. The p-value of all the slope terms is less than the alpha, so they are statistically significant
3. The model is statistically significant as the p-value for the model from F – the test is less than alpha.
4. Adj-R2 = 95.18% so 95.18%of the variation in the dependent variable is explained by the independent variable
5. The VIF of all the independent variables is less than 10, so there is no multicollinearity.

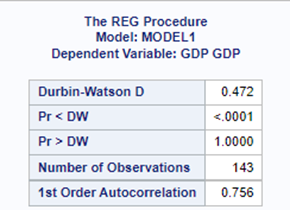
**Model Assumption**

****

1. According to the histogram, the residuals appear to be normally distributed, The assumption is true.

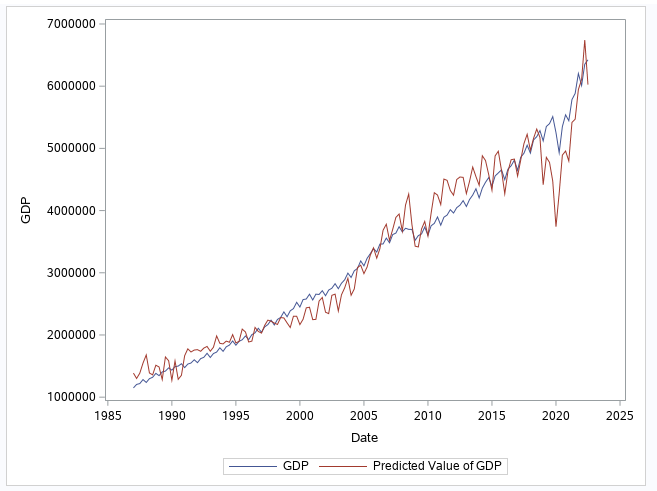
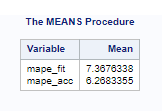


1. The residual vs. predicted value scatter plot does not show any pattern which indicates that the equal variances assumption is correct.

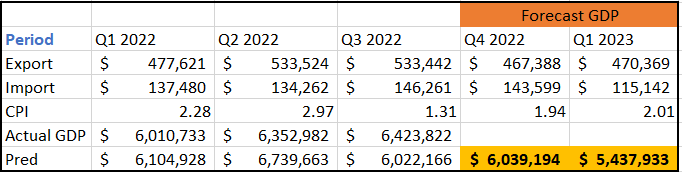
****

1. There is positive autocorrelation because the p-value is less than the alpha. The assumption is not true

**MAPE FIT and ACCURACY**



**FORECAST**



## 

## **Conclusion**

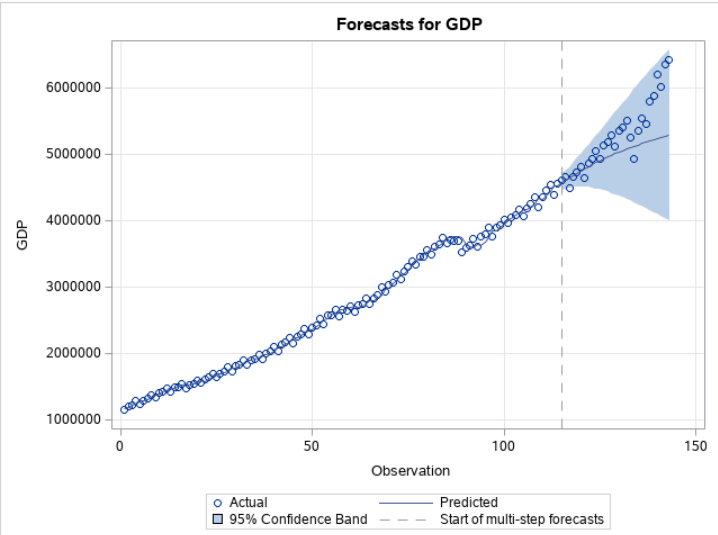
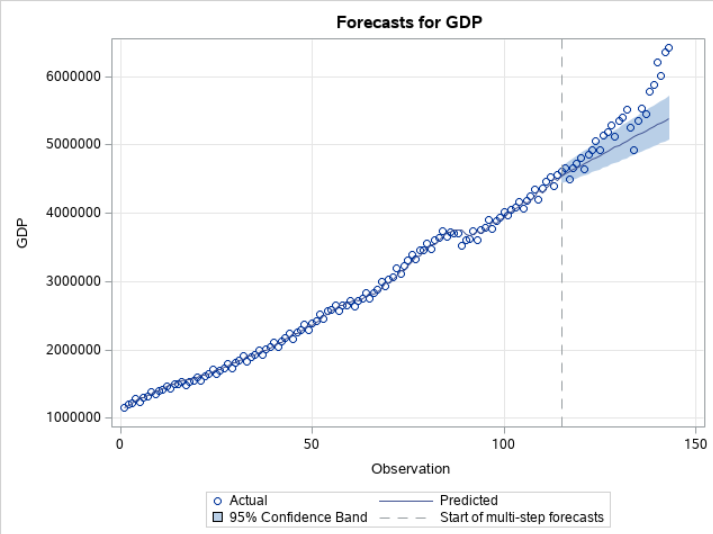
Comparing the MAPE Model FIT and Model Accuracy for models we have ran to forecast GDP:

|  | **Model** | **Model Fit** | **Model Accuracy** |
| --- | --- | --- | --- |
| **MAPE** | |
| **Scenario 1** | **Multiple Regression** | 7.36 | 6.26 |
| **Scenario 2** | **Holt’s** | 1.80 | 5.63 |
| **Damped** | 1.83 | 5.52 |
| **Linear Regression** | 3.94 | 7.45 |
| **ARIMA** | N/A | N/A |

Scenario 2 of our hypothesis proved true in that the Multiple regression model utilizing US Imports from China, US Exports, and CPI did not give us a better forecast than simpler traditional GDP forecasts based off of time.

Both Holt’s and Damped Exponential Smoothing models proved to be very good models to forecast for U.S. GDP. The difference between the two really comes down to whether US GDP will require a damped effect eventually or if it will continue to grow linearly forever.

**Damped Holts**

****

Damped’s confidence band includes the test set whereas Holt’s does not. However, the actual forecast line tapering off does not seem to be a correct prediction as US GDP still continues to rise and not top out any time soon.

## 

## 

## **SAS CODE (CPI)**

/\*Import data and analyze variable CPI\*/

proc import out=project datafile="/home/u62197416/sasuser.v94/Consolidated Data.xlsx"

dbms=xlsx replace;

run;

proc sgplot data=project;

series x=date y=cpi;

title "Quarterly US CPI % Change";

xaxis label="Quarter";

yaxis label="CPI (%QoQ)";

run;

proc timeseries data=project plots=acf out=\_null\_;

var cpi;

corr acf/nlag=48;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Simple Exponential Smoothing\*/

proc esm data=project print=all outfor=outproj lead=2 out=\_null\_ plot=forecasts;

forecast cpi/model=simple;

run;

proc sgplot data=outproj;

series x=\_timeid\_ y=actual;

series x=\_timeid\_ y=predict;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Creating 3 and 5 period moving average data points \*/

proc expand data=project out=project\_ma;

id date;

convert cpi=moving\_average3/transout=(movave 3);

convert cpi=moving\_average5/transout=(movave 5);

convert cpi=moving\_average8/transout=(movave 8);

run;

/\*Plotting new moving averages with original time series for comparison\*/

proc sgplot data=project\_ma;

series x=date y=cpi;

series x=date y=moving\_average3;

label moving\_average3="MA3";

series x=date y=moving\_average5;

label moving\_average5="MA5";

series x=date y=moving\_average8;

label moving\_average5="MA8";

run;

/\*Calculating MAPE of model fit which average is better?\*/

data project\_ma;

set project\_ma;

ape3=(abs(cpi-moving\_average3)/cpi)\*100;

ape5=(abs(cpi-moving\_average5)/cpi)\*100;

ape8=(abs(cpi-moving\_average8)/cpi)\*100;

run;

proc means data=project\_ma mean;

var ape3;

var ape5;

var ape8;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Winter’s Model \*/

/\* Multiplicative method \*/

proc esm data=project plot=forecasts lead=29 back=29 print=all outfor=outproject out=\_null\_;

id date interval=qtr;

forecast cpi/model=winters;

run;

/\* Additive method \*/

proc esm data=project plot=forecasts lead=29 back=29 print=all outfor=outproject1 out=\_null\_;

id date interval=qtr;

forecast cpi/model=addwinters;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*ARIMA\*/

proc arima data=project;

identify var=cpi(4) nlag=48 whitenoise=ignoremiss; /\*(4) does quarterly seasonal differencing is monthly\*/

estimate p=(3)(4)(8) q=(2) whitenoise=ignoremiss;/\*ARIMA(3,0,2)(2,1,0)4 () denotes seasonal in code\*/

estimate p=(4)(8) q=(1) whitenoise=ignoremiss;/\*ARIMA(0,0,1)(2,1,0)4 () denotes seasonal in code\*/

estimate p=(4) q=(1) whitenoise=ignoremiss;/\*ARIMA(0,0,1)(1,1,0)4 () denotes seasonal in code\*/

forecast id=date interval=qtr lead=2 out=out\_project;

run;

proc sgplot data=out\_project;

series x=date y=cpi;

series x=date y=forecast;

run;

## 

## **SAS CODE (EXPORTS)**

proc import out = export datafile="/home/u60891813/sasuser.v94/Consolidated Data\_Final.xlsx"

dbms= xlsx replace;

run;

proc sgplot data= export;

series x = date y = Exports;

title "Total Export ";

xaxis label = "Quarterely ";

yaxis label = "Export";

run;

PROC TIMESERIES DATA = export plots = acf out=\_null\_;

var Exports;

corr acf/nlag=48;

run;

/\* The data exibit the only Trend component So we choose between Holt's or Damped Trend model for model building\*/

proc esm data = export print=all lead=29 back=29 plot= forecasts out=\_null\_ outfor=ExportoutH ;

forecast Exports/model=linear;

run;

proc esm data = export print=all lead=29 back=29 plot= forecasts out=\_null\_ outfor=ExportoutD ;

forecast Exports/model=DAMPTREND;

run;

data export;

set export;

t=\_n\_;

export\_new=Exports;

if t > 114 then export\_new=.;

run;

/\* Linear fit\*/

proc reg data = export outest=export1;

model export\_new=t/aic bic adjrsq;

output out=Exports\_out1 r=Exports\_resid1 p= Exports\_pred1;

run;

data Exports\_out1;

set Exports\_out1;

mape\_fit = (abs(Exports\_resid1)/export\_new)\*100;

if t>114 then mape\_acc = (abs(Exports\_pred1-Exports)/Exports)\*100;

run;

proc means data = Exports\_out1 mean;

var mape\_fit mape\_acc;

run;

proc sgplot data = Exports\_out1;

series x =date y= Exports;

series x =date y= Exports\_pred1;

run;

/\*ARIMA\*/

proc arima data=export;

identify var=export\_new(1) nlag=36 whitenoise=ignoremiss;

estimate p=5 whitenoise=ignoremiss; /\* ARIMA(4,1,0) \*/

forecast id=date interval=qtr out=exportout;

run;

data export\_merge ;

merge exportout export;

run;

data Exports\_out2;

set export\_merge;

mape\_FIT = (abs(RESIDUAL)/export\_new)\*100;

if t>114 then mape\_acc = (abs(FORECAST-Exports)/Exports)\*100;

RUN;

proc means data = Exports\_out2 mean;

var mape\_FIT mape\_acc;

run;

proc sgplot data = exportout;

series x =date y= export\_new;

series x =date y= forecast;

run;

## **SAS CODE (IMPORTS)**

proc import out=project datafile="/home/u62188546/sasuser.v94/Project Files/Consolidated Data\_Final.xlsx"

dbms=xlsx replace;

run;

/\* the sgplot \*/

proc sgplot data=project;

series x=date y=imports;

run;

/\* ACF plot \*/

proc timeseries data=project plots=acf out=\_null\_;

var imports;

corr acf/nlag=52;

run;

/\* Multiplicative method \*/

proc esm data=project plot=forecasts lead=29 back=29 print=all outfor=outproject out=\_null\_;

id date interval=qtr;

forecast imports/model=winters;

run;

/\* Additive method \*/

proc esm data=project plot=forecasts lead=29 back=29 print=all outfor=outproject1 out=\_null\_;

id date interval=qtr;

forecast imports/model=addwinters;

run;

/\* Simple Regression Model \*/

proc reg data=project;

model imports=date/clb;

run;

data project;

set project;

t=\_n\_;

imports\_new=imports;

if t>114 then imports\_new=.;

run;

data project;

set project;

quarter=qtr(date);

if quarter=1 then q1=1; else q1=0;

if quarter=2 then q2=1; else q2=0;

if quarter=3 then q3=1; else q3=0;

if quarter=4 then q4=1; else q4=0;

run;

proc reg data=project outest=projectout;

model imports\_new=t q1 q2 q3/aic bic clb vif adjrsq dwprob;

output out=importsoutput p=pimports r=rimports;

run;

proc sgplot data=importsoutput;

series x=t y=imports\_new;

series x=t y=pimports;

run;

/\* MAPE \*/

proc reg data=importsoutput;

model imports\_new=t/clb;

output out=partout predicted=pimports residual=rimports;

run;

/\* Calculate MAPE \*/

data importsoutput1;

set importsoutput;

mapefit=(abs(rimports)/imports\_new)\*100;

if t>114 then mapeacc=(abs(pimports-imports)/imports)\*100;

run;

proc means data=importsoutput1 mean;

var mapefit mapeacc;

run;

/\* Seasonal ARIMA \*/

proc arima data=project;

identify var=imports(1,4) nlag=36 whitenoise=ignoremiss;

estimate p=(3)(4)(8) q=(2)(4) whitenoise=ignoremiss; /\* ARIMA(3,1,2)(2,1,1) \*/

\*estimate p=(2)(4)(8) q=(3)(4) whitenoise=ignoremiss; /\* ARIMA(2,1,3)(2,1,1) \*/

\*estimate p=(4)(4)(8) q=(3)(4) whitenoise=ignoremiss; /\* ARIMA (4,1,3) (2,1,1) \*/

forecast id=date interval=qtr lead=2 out=projectarimaout;

run;

data importmerge ;

merge projectarimaout project;

run;

data projectimport;

set importmerge;

mape\_fit = (abs(RESIDUAL)/imports\_new)\*100;

if t>114 then mape\_acc = (abs(FORECAST-imports)/imports)\*100;

RUN;

proc means data = projectimport mean;

var mape\_fit mape\_acc;

run;

proc sgplot data=projectimport;

series x=date y=imports;

series x=date y=forecast;

run;

## **SAS CODE (GDP)**

/\* GDP \*/

proc import out=usgdp datafile="/home/u62188546/sasuser.v94/Project Files/Consolidated Data\_Final.xlsx"

dbms=xlsx replace;

run;

/\* the sgplot \*/

proc sgplot data=usgdp;

series x=date y=GDP;

run;

/\* ACF plot \*/

proc timeseries data=usgdp plots=acf out=\_null\_;

var GDP;

corr acf/nlag=52;

run;

/\* Holt's Model \*/

proc esm data=usgdp print=all lead=29 back=29 plot= forecasts out=\_null\_ outfor=usgdpout1 ;

forecast GDP/model=linear;

run;

/\* Damped Model \*/

proc esm data=usgdp print=all lead=29 back=29 plot= forecasts out=\_null\_ outfor=usgdpout2 ;

forecast GDP/model=damptrend;

run;

/\* Simple Linear Regression Model \*/

data usgdp;

set usgdp;

t=\_n\_;

GDP\_new=GDP;

if t>114 then GDP\_new=.;

run;

proc reg data=usgdp outest=usgdp1;

model GDP\_new=t/aic bic clb vif adjrsq dwprob;

output out=GDPout1 r=GDPr p= GDPp;

run;

data GDPout1;

set GDPout1;

mape\_fit=(abs(GDPr)/GDP\_new)\*100;

if t>114 then mape\_acc=(abs(GDPp-GDP)/GDP)\*100;

run;

proc means data=GDPout1 mean;

var mape\_fit mape\_acc;

run;

proc sgplot data=GDPout1;

series x=date y=GDP;

series x=date y=GDPp;

run;

/\*ARIMA - GDP\*/

data project;

set project;

t=\_n\_;

gdp\_new=gdp;

if t > 114 then gdp\_new=.;

run;

proc arima data=project;

identify var=gdp\_new(1) nlag=36 whitenoise=ignoremiss;

estimate p=3 whitenoise=ignoremiss; /\* ARIMA(3,1,0) \*/

forecast id=date interval=qtr out=projectout;

run;

data project\_merge ;

merge projectout export;

run;

data project\_out2;

set project\_merge;

mape\_FIT = (abs(RESIDUAL)/gdp\_new)\*100;

if t>114 then mape\_acc = (abs(FORECAST-gdp)/gdp)\*100;

RUN;

proc means data = project\_out2 mean;

var mape\_FIT mape\_acc;

run;

proc sgplot data = projectout;

series x =date y= export\_new;

series x =date y= forecast;

run;

/\*Multiple regression GDP\*/

proc sgplot data= Export;

series x=date y= GDP;

run;

PROC TIMESERIES DATA = export plots = acf out=\_null\_;

var GDP;

corr acf/nlag=48;

run;

Data gdp\_pred;

set export;

interraction\_term1 = CPI\*IMPORTS;

interraction\_term2 = CPI\*EXPORTS;

RUN;

proc reg data=EXPORT outest=gdp1;

model GDP = exportS CPI Imports/clb corrb vif aic bic adjrsq;

output out=GDPoutput p=gdp\_pred r=gdp\_resid;

run;

proc reg data=gdp\_pred outest=gdp3;

model GDP = Exports CPI interraction\_term1/clb corrb vif aic bic adjrsq;

output out=GDPoutput p=gdp\_pred r=gdp\_resid;

run;

proc reg data=gdp\_pred outest=gdp2;

model GDP = IMPORTS CPI interraction\_term2/clb corrb vif aic bic adjrsq dwprob;

output out=GDPoutput p=gdp\_pred r=gdp\_resid;

run;