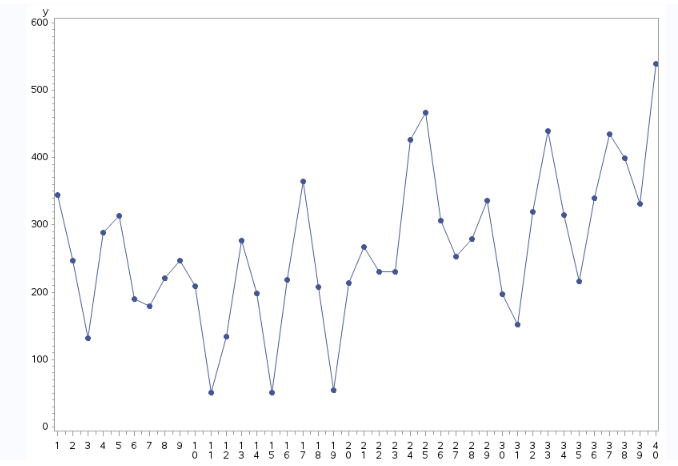
**Predictive Analytics and Business Forecasting**

Anannya Chatterjee – Exam 2

Answer 1: The answer is in SAS code.

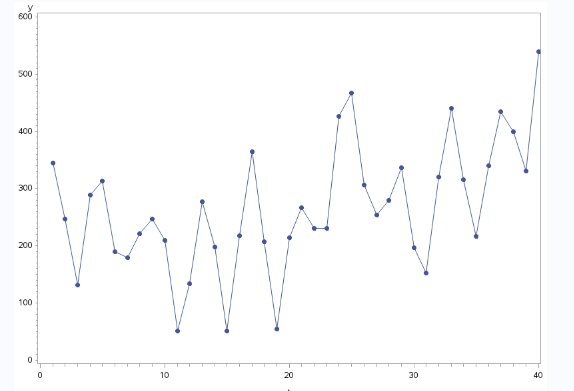
**Answer 2:** Here I am solving question 6.4 from textbook

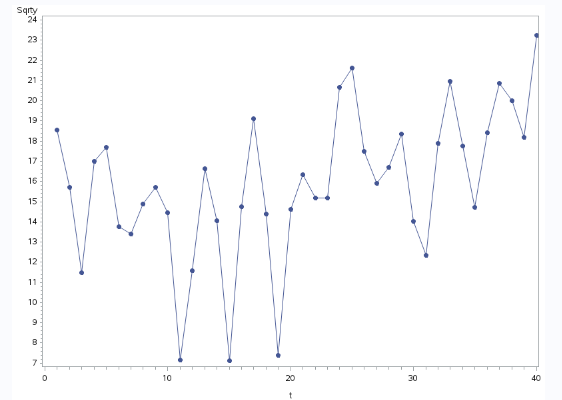
I have plotted the time series.

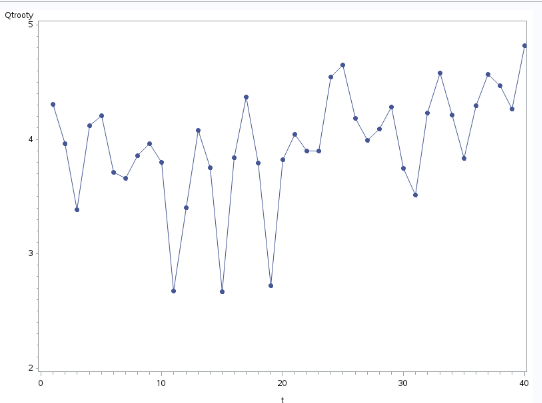


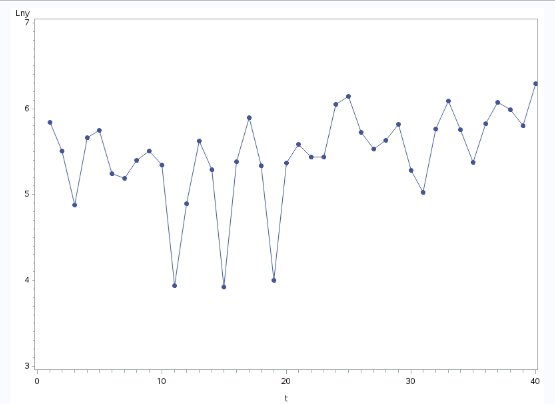
1. Based on the plot, I can see that there is a quadratic trend. As time is increasing, energy bills sometimes increase and then again decrease.
2. Based on the plot, it shows here increasing seasonal variation. It is not constant. And I have tried to transform it.

I have done the transformations within a data step. I have plot each of y, square root of y, quartic root of y, and natural logarithm of y versus Time. The natural logarithm of y versus Time gives a constant seasonal variation as shown below.









1. The Dummy variables ( Q1, Q2, Q3) are defined as:

data ebill3;

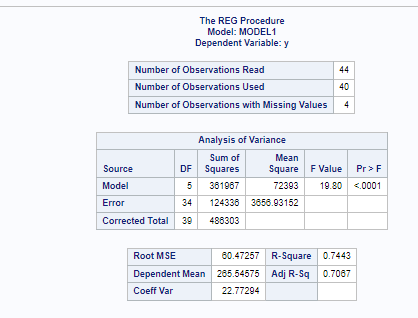
set energybill;

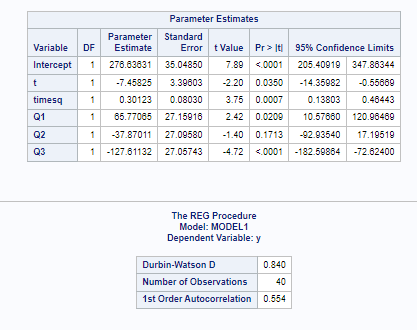
if mod(t,4)=1 then Q1=1; else Q1=0;

if mod(t,4)=2 then Q2=1; else Q2=0;

if mod(t,4)=3 then Q3=1; else Q3=0;

timesq=t\*\*2;





When we see the result, the intercept, t, timesq, Q1 and Q3 and statistically significant because the p value is less then alpha which is 0.05.

But Q2 is statistically not important because the p value is 0.1713 which is more than alpha (0.05).

1. The prediction equation for the model is:

**Y = 276.636 - 7.458\*t + 0.301\*timesq + 65.77\*Q1 – 37.87\*Q2 - 127.61\*Q3**

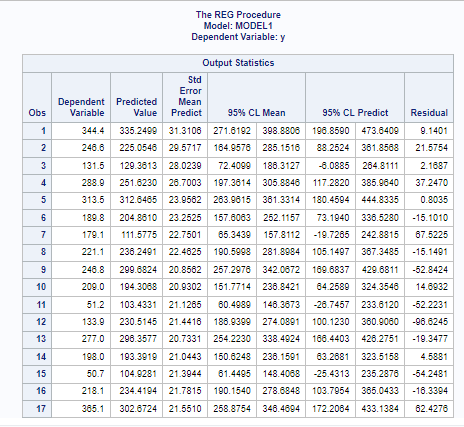
**For finding out Y(hat) 41, I replaced t = 41, timesq = 1681, Q1 = 1 AND Q2=Q3=0**

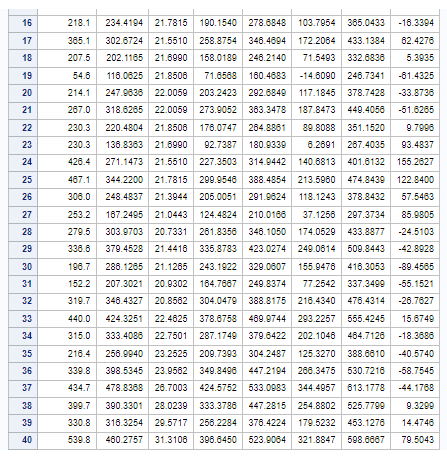
**Y(hat)41 = 542.609**

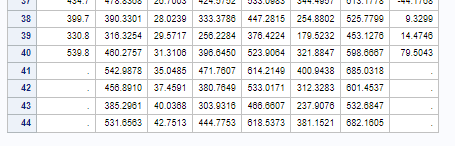
**For finding out Y(hat) 42, I replaced t = 42, timesq = 1764, Q2 = 1 AND Q1=Q3=0**

**Y(hat)42 = 456.494**

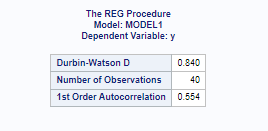
1. Report point forecasts and 95% prediction intervals for the energy bills in periods 41,42, 43, and 44 are show below.







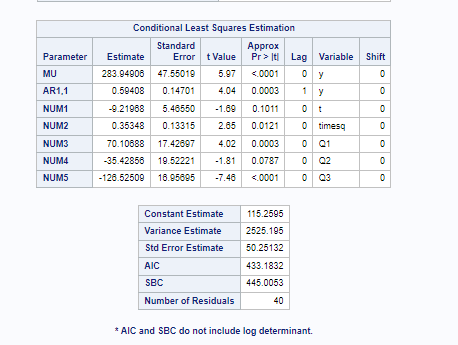
1. The Durbin – Watson test is conducted and the value is **0.840** which is from 0 to less than 2 and that indicates positive autocorrelation. It is shown below.

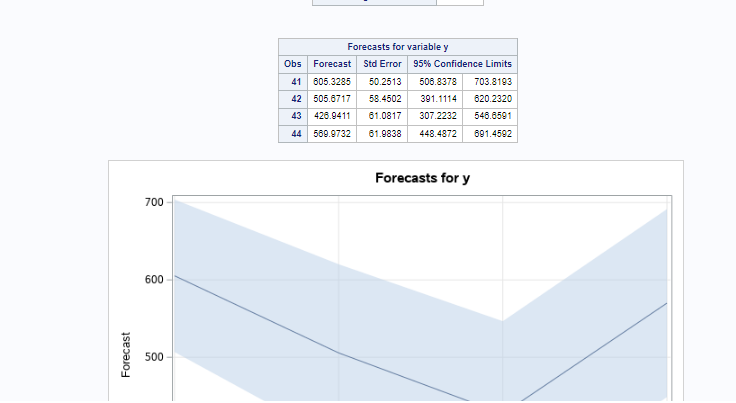


1. Hence the independency assumption is getting violated and so to improve the regression model, we must use some kind of autoregressive process.

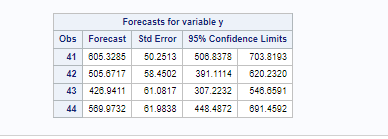
We will try to implement ARIMA model for this dataset now.

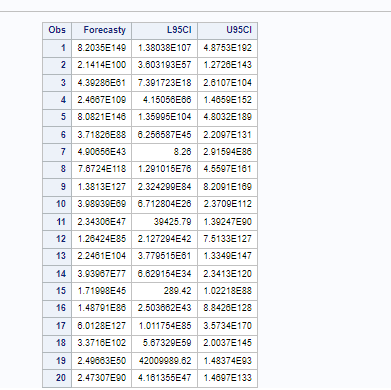
The value of ‘phi’ is statistically significant as it is less than alpha(0.05). Also other independent variables seem important other than the Q2 (because the value of p is more than alpha for Q2).

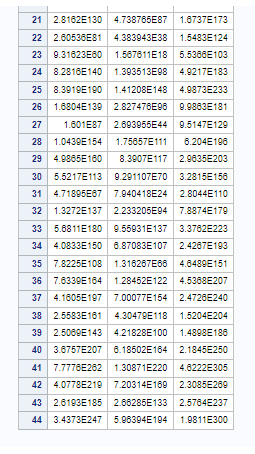




3. The point forecasts and 95% prediction intervals for the energy bills in periods 41,42, 43, and 44 is shown below.





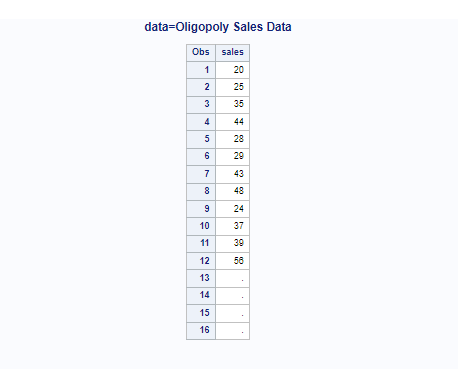


**Answer 3:**

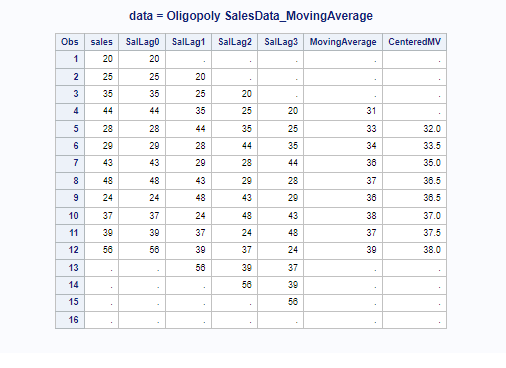
Book Question 7.2

A and b:

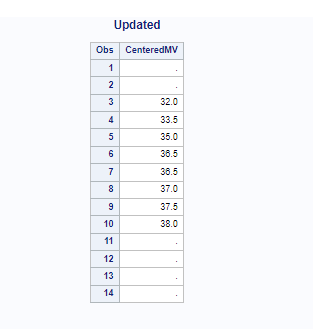
The data is shown by SAS in a tabular format.



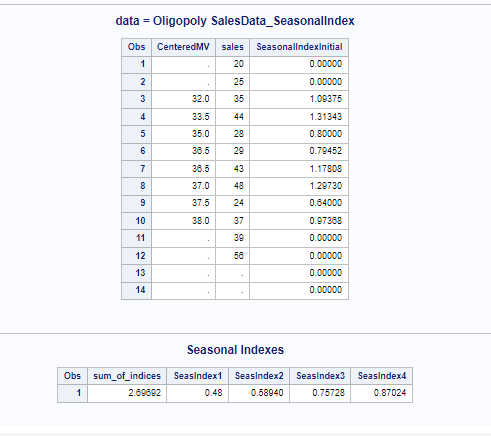
The moving average and centered moving averages are calculated as shown below.



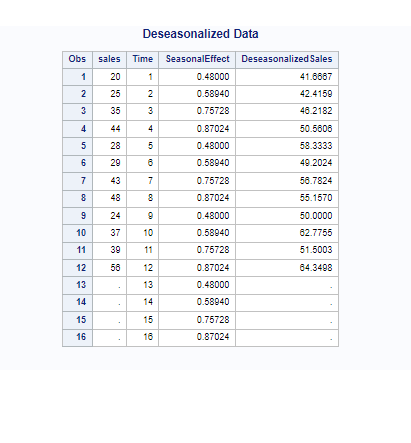
c.



d.



e.



f.The deseasonalized observations are plot versus time. A linear trend exists.

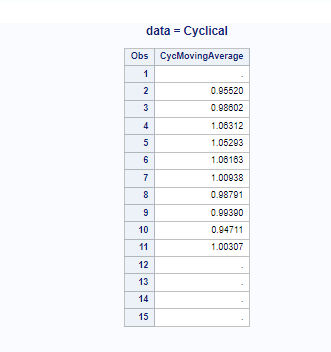


g. A linear trend appears. From the result, B0 = 42.35 and B1 = 1.54693

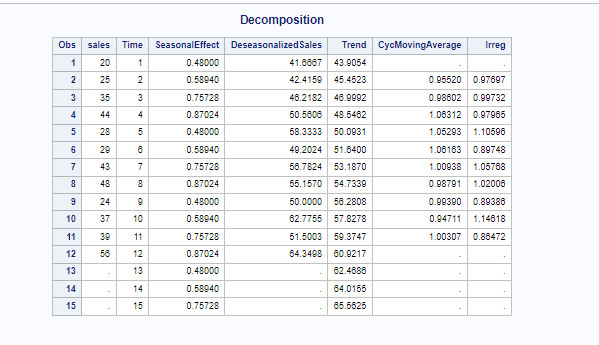
The trend line beomes:

TR = 42.35 + 1.54\*time

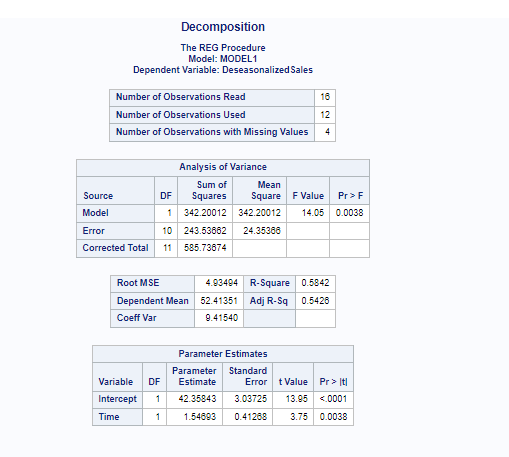
h.



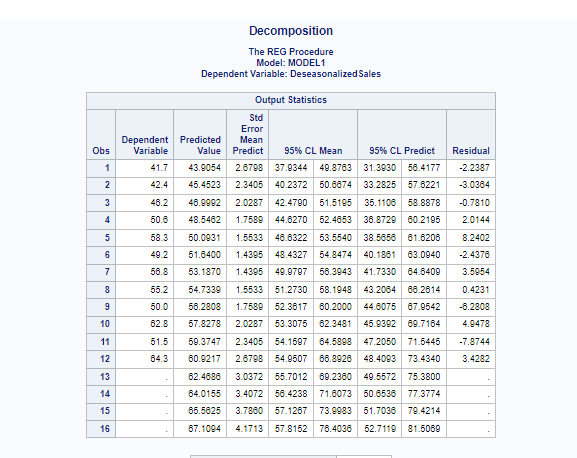
I



j.



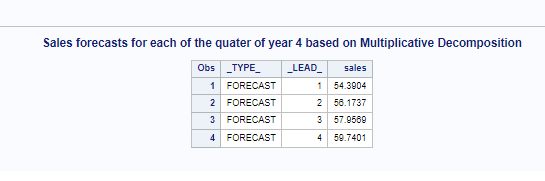
k.



l. The cl values define well defined cycle when it is seen in the result.

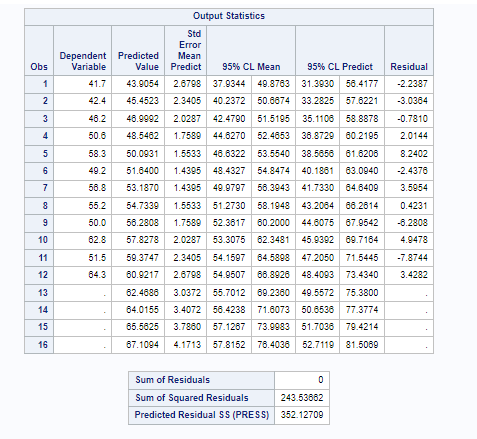
m.

Point forecasts of Oligopoly sales for each quarter of year 4 is computed in SAS and shown below.



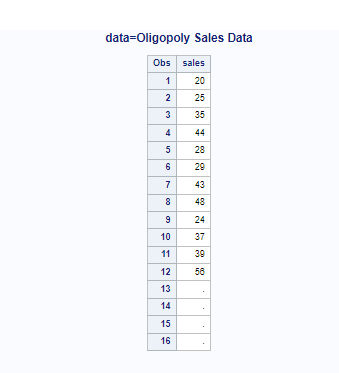
n.

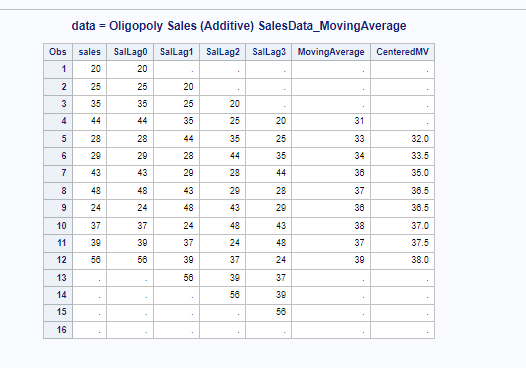
Approximate 95% prediction interval forecasts of Oligopoly sales for each quarter of year 4 is shown below after calculating in SAS.

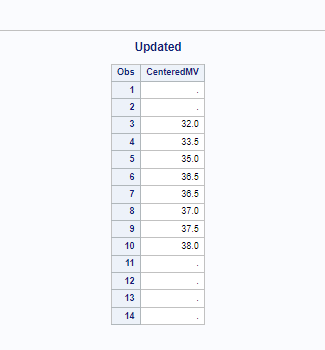


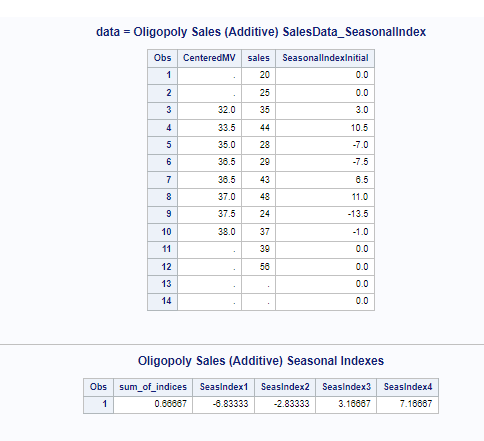
Book Question 7.4

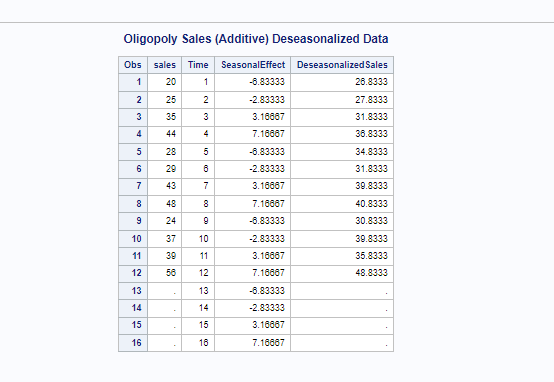
The whole data is now analyzed using additive decomposition.

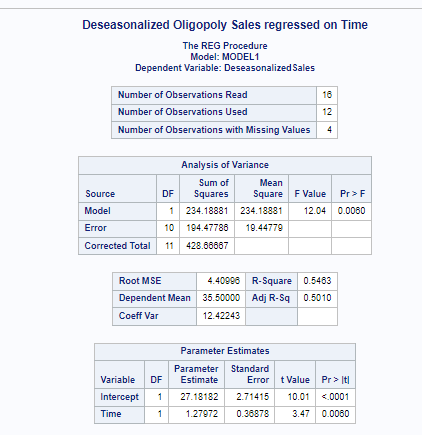


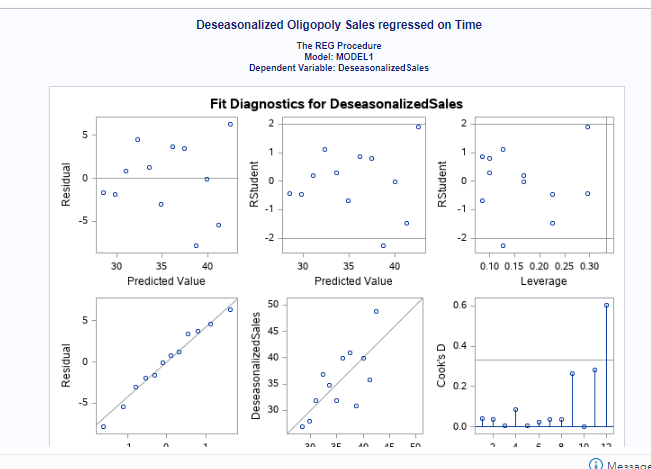




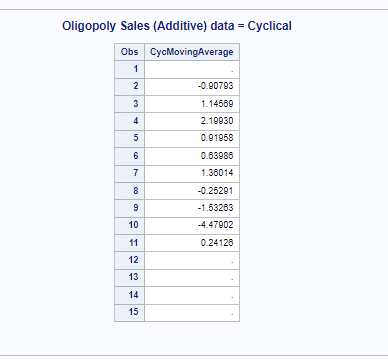


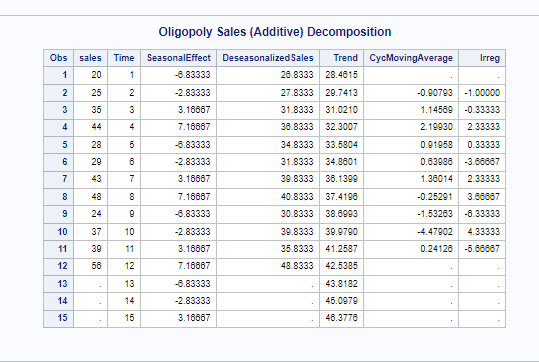


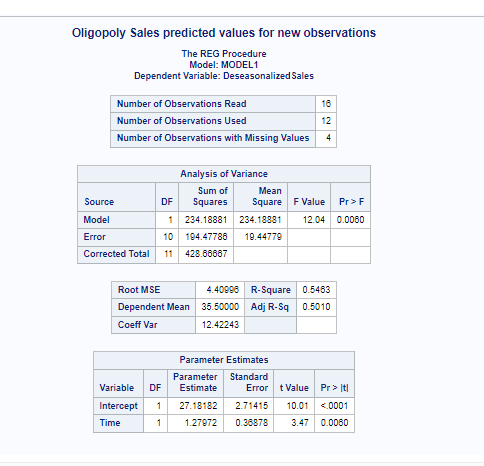


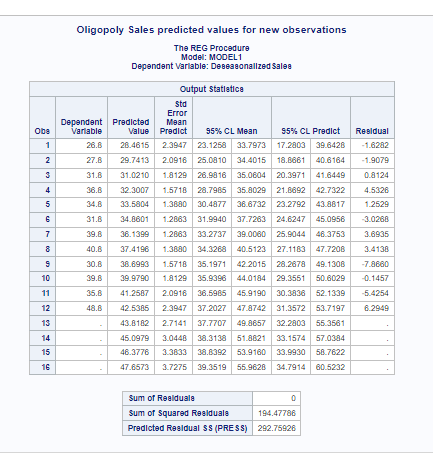


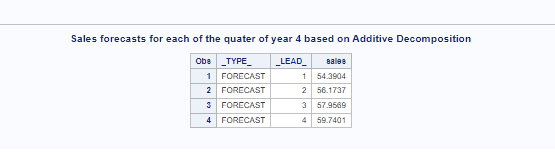












After comparing the multiplicative method and the additive method, I come to an understanding that both are equally beneficial for this dataset and they both give the same predicted value for each of the quarter for year 4. So, both are appropriate.

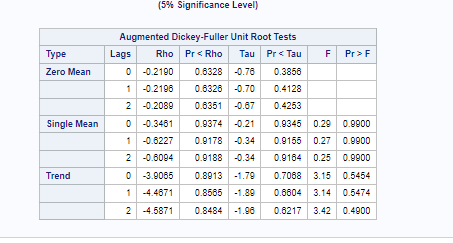
**Answer 5:**

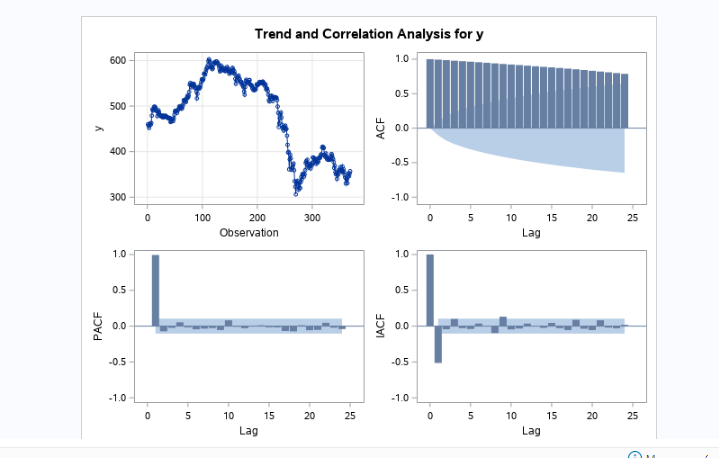
1. The Cryptocurrency daily stock prices when plotted do not seem stationary. I have used Augmented Dickey-Fuller Unit Root Tests to check about the stationarity of the data using the below command.

proc arima data=work.stockprice; /\*PROC ARIMA\*/

identify var=y scan stationarity=(adf); /\*Generate SAC and SPAC for y\_t\*/

As we see below, the p value is much more than alpha which is 0.05 in the ADF root test.





Hence it is confirmed that the cryptocurrency daily stock prices data is not stationary.

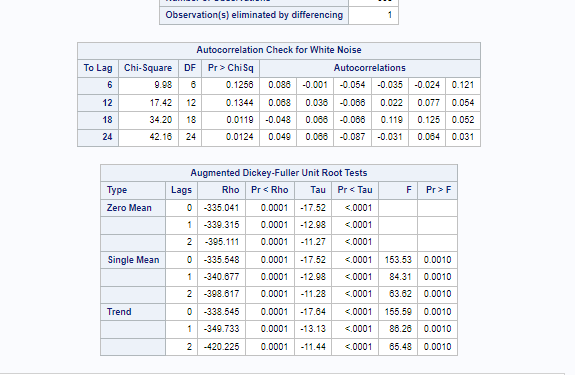
Now I will **try First differencing** for the stock data using the below command.

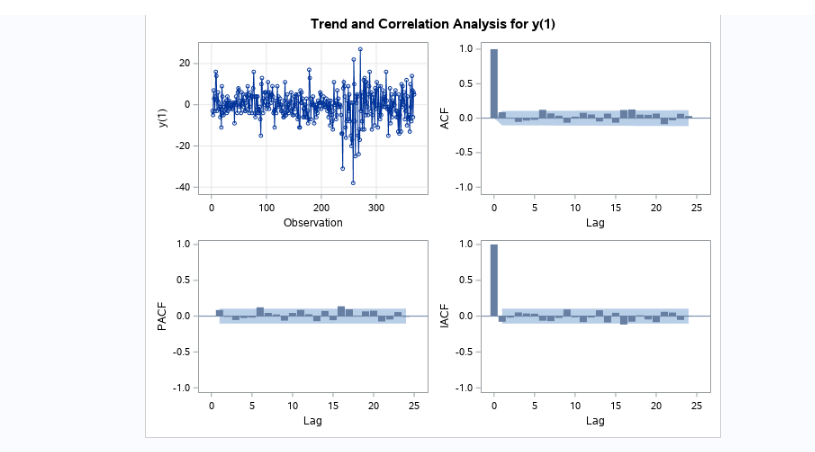
**identify var=y(1) stationarity=(adf);** /\*Generate SAC and SPAC for z\_t\*/

After running this, I confirm that now the Daily stock prices data is stationary. That is also confirmed from the Dicker Fuller test results. All the p value is much lesser than alpha which is 0.05 in the ADF root test.

Also multicollinearity do not exist.

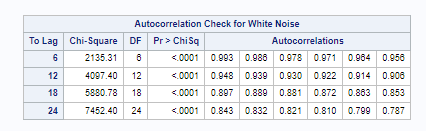
The below picture shows that.



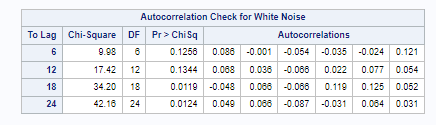


5b.

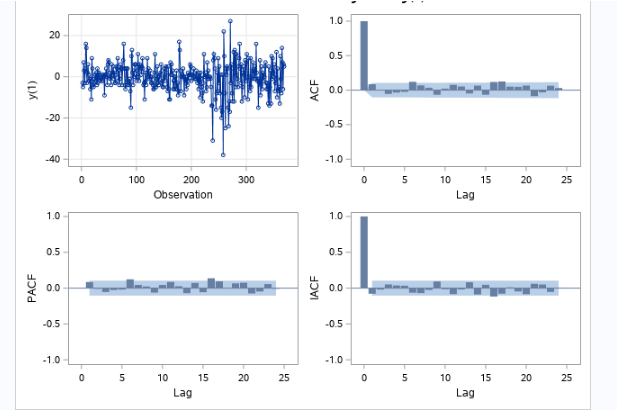
As we get from the Ljung Box test, the sample auto correlation of the original time series is very less than alpha which is 0.05. Here the p value is lesser than 0.0001.



But when I transformed the stock price time series using First differencing, I got better result in the Ljung box text for autocorrelation better p value which is higher than alpha and so the model is adequate.



5c. Now I am describing the behavior of SAC and SPAC in this first differencing is completed and the transformed time series is stationary.



We can see here that SAC dies down very quickly and SPAC cuts off after lag 6 which will help us to build AR model.

Also if we investigate more we can think SPAC dies down very quickly and SAC cuts off after lag 6 which will help us to build MA model.

1. Let us try building models now.

**Model 1:**

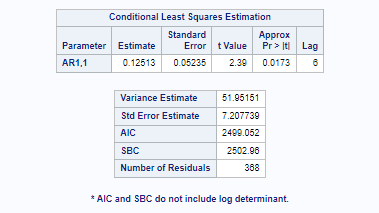
SAC dies down very quickly

SPAC: SPAC cuts off after lag 6.

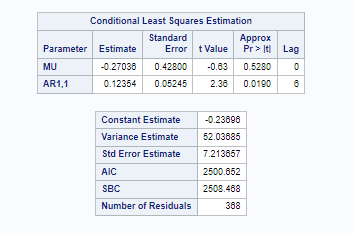
So, we can estimate an Auto regressive model of order 1.

This is AR(1) . Here , 1 is lag 6.

AR(1) with “No Constant”



AR(1) with “Constant”:



So whever we compare “No constant” and “with constant “, we find here that Constant is not required for this data.

As we see p value of constant is 0.528 which is much higher than 0.05 alpha. Which makes it statistically insignificant. Also, Standard Error estimate and AIC increases when constant is there.

**So, the model doesn’t need Constant.**

Model 2:

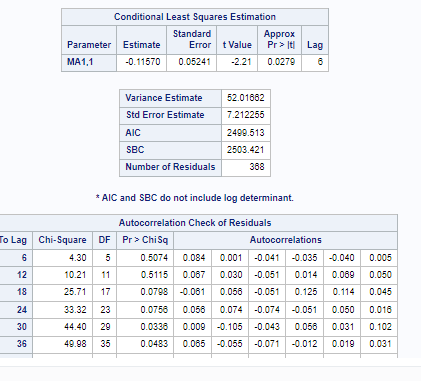
SAC cuts off after lag 6.

SPAC dies down very quickly

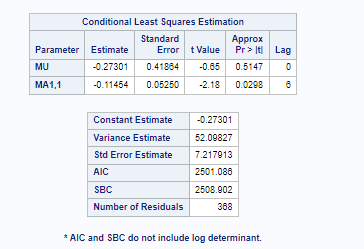
So, we can estimate a Moving Average model of order 1.

This is MA(1) . Here, 1 is lag 6.

Without Constant:



With Constant:



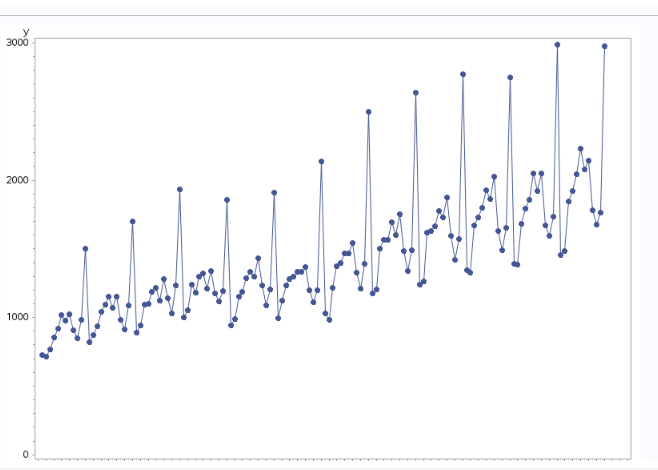
So overall checking all the results**, AR model with order 1 without Constant is the best fit model.** We cannot also any spikes in the residual correlation also.

**Answer:6**

**11.10**

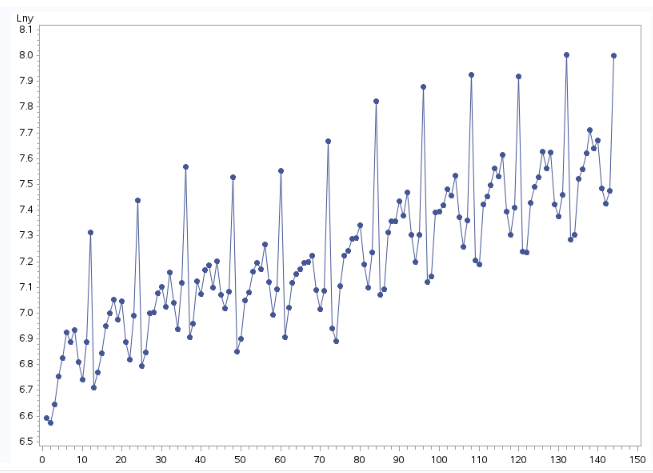
a. I have written the SAS code for plotting the Logarithmic transformation of the Y variable in the sports good’s dataset.



The original time series plot and logarithmic transformed time series plot are shown here.

As we see the original time series has increasing variance. But after I transform it to logarithmic time series, still it shows increasing variance.

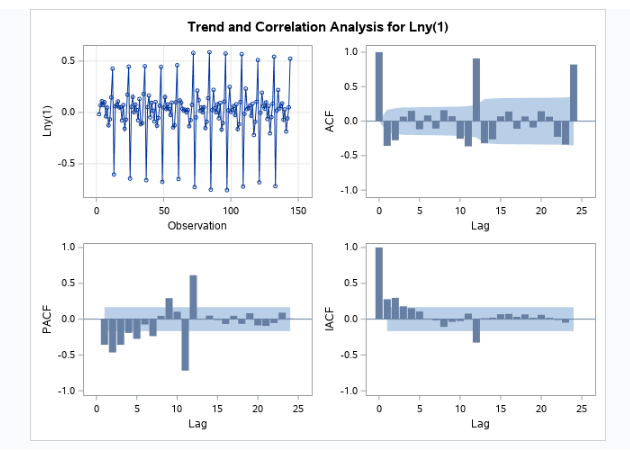
This is shown below.



1. The values produced by the transformation that is mentioned below might have constant variance. But for confirming that , we need to check in SAS.



1. I will now plot the trend and correlation analysis for Lny(1) as shown below.

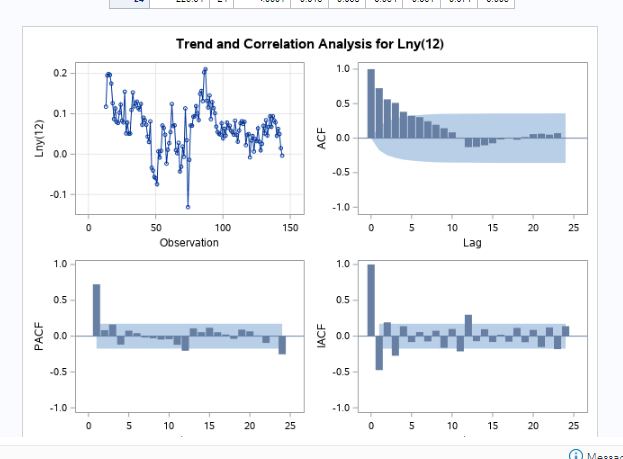


So, I see here that the SAC dies down extremely slowly at the seasonal levels, so the logarithmic transformation at seasonal levels is not stationary.

Hence, **Logarithmic transformation at first difference is not stationary.**

Now we will go for **seasonal differencing.**

As depicted below, we see the seasonal differencing of Lny time series.



If we see here, at seasonal levels, SAC dies down pretty fast and SPAC has spike at 1 and 3 and SPAC cuts off after lag 3.

So, we can create a Auto Regressive model with order 2. (lags at 1 and 3)

So, at non seasonal levels, we come up with



Now discussing about the Seasonal level, we have

SAC dies down at seasonal level and SPAC cuts off after lag 12. So we come up with AR model with order 1.

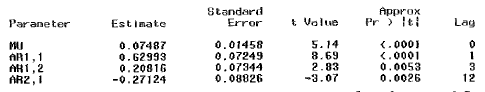


And now when combine these facts, we get



**11.11**

a. From the figure 11.31, we see that the least square point estimates of the parameters are:



Where the parameter AR1,1 corresponds to phi1,

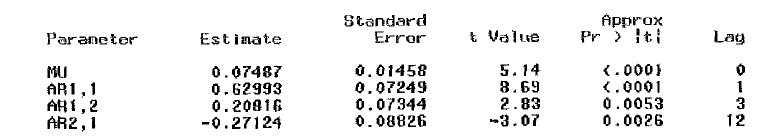
AR1,2 corresponds to phi3,

AR2,1 corresponds to phi1,12 and MU corresponds to 

So, delta = 0.07487, phi1 = 0.62993, phi2 = 0.20816 and phi1,12 = -0.27124

b.

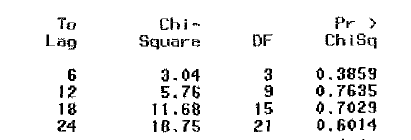
Since we know that a parameter is significant in model and has to be kept if the t value is significant. And t value is significant if the absolute t value is higher or equal to 1.96, meaning |t|≥1.96.



So here, mod of t value is greater than 1.96 for all the parameters.

So, all the parameters should be kept in the model.

C.



From this data that is given regarding the autocorrelation check of residuals, we can find out the p value for a specific lag.

P6 = 0.3859

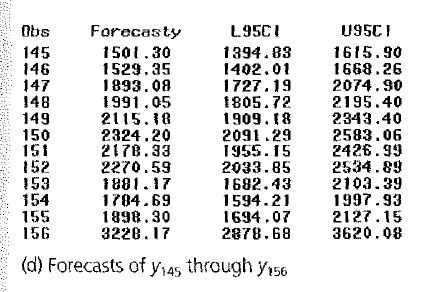
P12 = 0.7635

P18 = 0.7029

P24 = 0.6014

Now if we compare these p values to the alpha = 0.05, we see that all are greater than 0.05 and hence we cannot reject the adequacy of the model. The model is adequate.

1. We are given the forecasts value of Y.



From here we can find that the **point forecast for y145 is 1501.30.**

**95% prediction interval of y145** spans from

**L95CI: 1394.83**

**U95CI: 1615.90**

**11.15**

1. Here if we see at the SAS output figure, at non seasonal levels, we can find out SPAC cuts off after lag 3 and SAC dies down fast.

So, we can build an AR model with order 2 because the spike is at lag1 and lag 3. Hence the equation we can build for AR model here is:



Similarly, at seasonal levels, SAC dies down fast and SPAC cuts off after lag 12.

So, we can build an AR model with order 1.

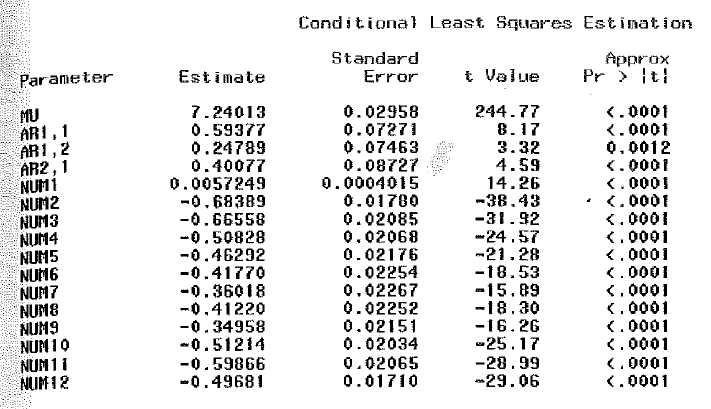
This gives us the equation:



When we combine, we get the reasonable choice for a tentative model.

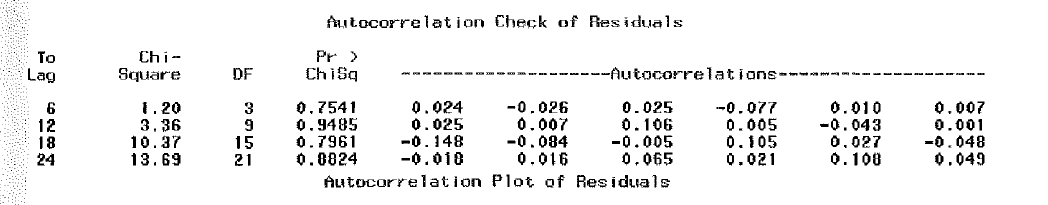
1. Looking at the Figure 11.37, we see the various SAS outputs of the regression model.

All the parameters look statistically significant as shown below. The p value is much lesser than alpha. All the p values are lesser than 0.05.



Also, we see that the model is adequate. The Ljung Box test says that if the p value is pore than 0.05b(alpha), we reject to consider it inadequate.

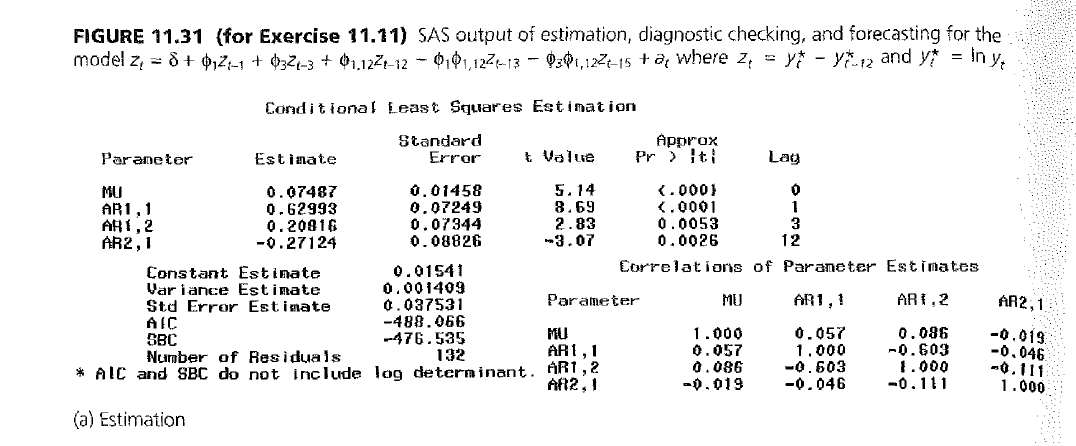
As shown below, all the p values are more than 0.05,



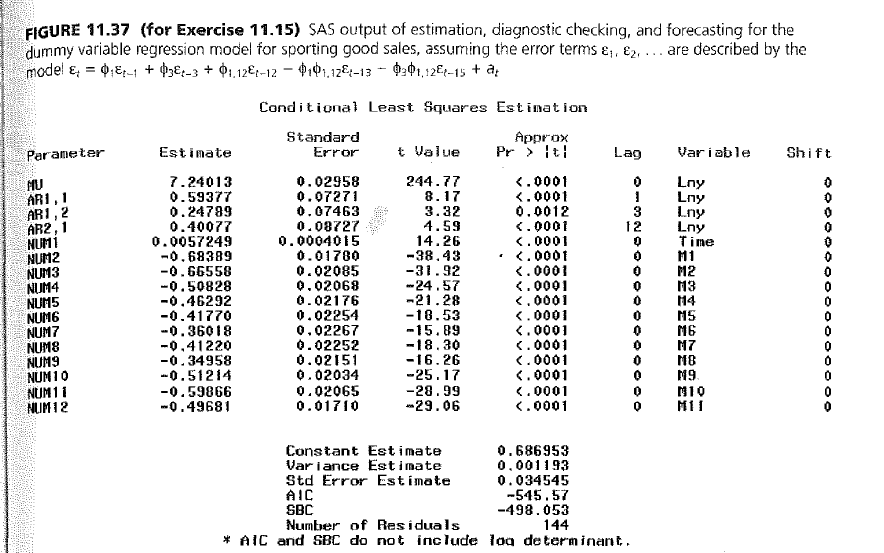
Also, there is no multicollinearity between the independent variables.

So, I find that this is a reasonable model to forecast retail sales.

1. Now when I compare the two models as shown below:



Here the standard error estimate is **0.037531**



Here the standard error estimate is **0.034545**

**Hence the** standard error is more in the model in Figure 11.31.

--------------------XXXXXXX-----------------------------